ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Recent Advances in Engineering Systems

2018-19

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 *Recent Advances in Engineering Systems (2018–2019)* in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* presents a timely and comprehensive collection of research contributions that reflect the continued evolution of engineering systems in an era defined by rapid technological progress and increasing system complexity. The period of 2018–2019 has been marked by significant strides in the development of integrated, intelligent, and adaptive systems capable of addressing the multifaceted challenges of modern society. This special issue brings together innovative studies that span theoretical advancements, methodological improvements, and practical implementations across a wide range of engineering domains.

A central theme across the contributions is the growing integration of digital technologies within engineering systems. The emergence of data-driven approaches, artificial intelligence, and advanced simulation techniques has transformed how systems are designed, analyzed, and managed. Researchers have increasingly leveraged machine learning algorithms, real-time data acquisition, and predictive modeling to enhance system efficiency, reliability, and responsiveness. These advancements are particularly evident in applications such as smart infrastructure, autonomous systems, industrial automation, and intelligent transportation, where the ability to process and respond to complex data streams is essential.

Another important focus of this special issue is the advancement of system optimization and performance enhancement. The included studies explore novel optimization techniques, control strategies, and decision-making frameworks that aim to improve system functionality under varying and often uncertain conditions. By incorporating robust design principles and adaptive control mechanisms, these contributions demonstrate how engineering systems can achieve higher levels of performance while maintaining stability and resilience. Such approaches are crucial in ensuring that modern systems can operate effectively in dynamic environments.

Sustainability and resilience continue to play a significant role in shaping contemporary engineering research. The works presented in this issue emphasize the need for systems that are not only efficient but also environmentally responsible and capable of withstanding external stresses. From renewable energy integration and sustainable construction practices to resource-efficient system design, the research highlights innovative solutions that align engineering advancements with global sustainability objectives. These efforts underscore the importance of balancing technological progress with environmental stewardship.

Interdisciplinary collaboration remains a defining characteristic of the advancements showcased in this issue. The complexity of modern engineering challenges necessitates the integration of knowledge from multiple fields, including civil, mechanical, electrical, and computer engineering, as well as applied sciences. The contributions illustrate how such collaboration fosters innovation, enabling the development of comprehensive systems that are more adaptable, efficient, and robust. This interdisciplinary approach is particularly evident in the design and implementation of cyber-physical systems, where physical processes are closely integrated with computational intelligence.

The editorial team expresses its sincere gratitude to the authors, reviewers, and editorial members whose dedication and expertise have contributed to the successful completion of this special issue. Their efforts have ensured the publication of high-quality research that advances

both theoretical understanding and practical application in engineering systems. The collection presented herein reflects the ongoing transformation of engineering systems into more intelligent, interconnected, and sustainable entities, providing valuable insights and inspiration for future research and development in this rapidly evolving field.

Guest Editor

Prof. Ahmad Yusairi Bani Hashim

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **July 2018**

CONTENTS

Automation System for Regulation Optimization in Power Transformer Design
by Tatjana Šimović and Mislav Gazdović

Robust synchronization of nonfragile control of complex dynamical network with stochastic coupling and time-varying delays
by Fehrs Adu-Gyamfi, Yuhua Cheng and Chun Yin

A Wearable Exoskeleton Rehabilitation Device for Paralysis – A Comprehensive Study
by Ahmed Roshdy, Samer Al Kork, Sherif Said and Taha. Beyrouthy

High-Temperature Optical Characterization of Wide Band Gap Light Emitting Diodes and Photodiodes for Future Power Module Application
by Syam Madhusoodhanan, Abbas Sabbar, Sattar Al-Kabi, Stanley Atcitty, Robert Kaplar, Binzhong Dong, Jiangbo Wang, Shui-Qing Yu and Zhong Chen

Design of an Additively Manufacturable Multi-Material Light-Weight Gripper with integrated Bellows Actuators
by Gabriel Dämmer, Sven Gablenz, Alexander Hildebrandt and Zoltan Major

A Practical PIR-based Scheme for Discovering Nearby Places for Smartphone Applications
by Maryam Hezaveh and Carlisle Adams

Stress Level Classification Using Heart Rate Variability
by Tlija Amira, Istrate Dan, Badii Atta, Gattoufi Said, Bennani Az-eddine and Wegrzyn-Wolska Katarzyna

A Proposed Architecture for Parallel HPC-based Resource Management System for Big Data Applications
by Waleed Al Shehri, Maher Khemakhem, Abdullah Basuhail and Fathy E. Eassa

Comparison of Electrogastrograms in Seated and Supine Positions Using Wayland Algorithm
by Yasuyuki Matsuura and Hiroki Takada

Morphological and Optoelectrical Characterization of Silicon Nanostructures for Photovoltaic Applications
by Babacar Dieng, Moussa Toure, Modou Beye, Diouma Kobor and Amadou Seidou Maiga

W.A.S.T.E. R.E.D.U.C.E.: Waste Auditing Sensor Technology to Enhance the Reduction of Edible Discards in University Cafeterias & Eateries
by Shiree Hughes, Jiannan Zhai and Jason Hallstrom

A Relation Extraction System for Indian Languages
by Ajees Arimbassery Pareed and Sumam Mary Idicula

A Spectrum Sharing based Metering Infrastructure for Smart Grid Utilizing LTE and WiFi
by Imtiaz Parvez and Arif I. Sarwat

State Estimation based Echolocation Bionics and Image Processing based Target Pattern Recognition
by David Kondru, Mehmet Celenk and Xiaoping A. Shen

LabVIEW Development for an Intelligent Management System of the Electrical Energy Free Market
by Rodrigo Reina Muñoz, Rafael Valim Xavier de Souza and Alex Soto da Silva

Web Authentication: no Password; Listen and Touch
by Viorel Lupu

The Visualization of Cattle Movement Data in The State of Pará in 2016 Through Networks of Animal Transit Graphs and Guides
by Samuel Carvalho de Aragão, Agnaldo Reis Pontes, Luis Manuel Borges Gouveia, Samuel Franco Lopes, Pier Kenji Rauschkolb Katsuda, Anirene Galvão Tavares Pereira, Márcio Teixeira Oliveira, Jefferson Pinto de Oliveira, Rita do Socorro Brito Coroa, Gilson Ferreira Araújo and Marcio Merêncio Panza de Siqueira

Optimal Designs of Constrained Accelerated Life Testing Experiments for Proportional Hazards Models
by Xiaojian Xu and Wanyi Huang

Assessment of Coordinated Multipoint Transmission Modes for Indoor and Outdoor Users at 28 GHz in Urban Macrocellular Environment
by Muhammad Usman Sheikh, Ritayan Biswas, Jukka Lempiainen and Riku Jantti

Conducted and Radiated Interference on Interconnection's Lines
by Patricio E. Munhoz-Rojas

Transfer Function Analysis of Fractional-Order Three-Dimensional Electrically Coupled Cell Network
by Mahmut ÜN

Frequency-Based Design of Electric System for Off-shore Wind Power Plant (OWPP)
by Uttam S. Satpute, Diwakar R. Joshi and Shruti Gunaga

Low Contrast Image Enhancement Using Convolutional Neural Network with Simple Reflection Model
by Bok Gyu Han, Hyeon Seok Yang, Ho Gyeong Lee and Young Shik Moon

Photodecoloration of Methyl Orange Solution Assisted by ZrS3 Powders
by Sofya Artemkina, Anastassiia Poltarak, Pavel Poltarak, Igor Asanov and Vladimir Fedorov

Development & Implementation of Smart Vehicle Over Speeding Detector using IoT

by Mohammad Ahmar Khan, Sarfraz Fayaz Khan, Taher Alawi Taher Barham BaOmar and Abdul Rahman Mohammed Aqeel Ba Omar

A Novel Pulse Position Modulator for Compressive Data Acquisition
by Constantine A. Pappas

Enhancing and Monitoring Patient Outcomes Through Customized Learning
by Majed Almotairi, Mohammed Abdulkareem Alyami and Yeong-Tae Song

Nonlinear Analytic Modeling for Novel Linear Variable Reluctance Motors
by Aymen Lachheb, Jalel Khediri and Lilia El Amraoui

Digitization Impact Assessment Model for Secondary Schools: Case of Nairobi County in Kenya
by Nicodemus Kalugho Mwambela and Simon Nyaga Mwendia

Indoor Positioning and Path Planning Platform for iRobot Create 2 Sweeping Robot
by Rung-Shiang Cheng, Wei-Jun Hong, Jhe-Lin Li and Kawuu W. Lin

Development of Application Specific Electronic Nose for Monitoring the Atmospheric Hazards in Confined Space
by Muhammad Aizat Bin Abu Bakar, Abu Hassan Bin Abdullah and Fathinul Syahir Bin Ahmad Sa'ad

Experimental Investigation of Human Gait Recognition Database using Wearable Sensors
by Sherif Said, Samer Al-Kork, Vishnu Nair, Itta Gowthami, Taha Beyrouthy, Xavier Savatier and M Fayek Abdrabbo

iSensA – A System for Collecting and Integrating Sensor Data
by João Manuel Leitão Pires Caldeira, Vasco Nuno da Gama de Jesus Soares, Pedro Miguel de Figueiredo Dinis Oliveira Gaspar, Joel José Puga Coelho Rodrigues, Ricardo Manuel Valentim Fontes and José Luís Lopes Silva

Trajectory Tracking Control Optimization with Neural Network for Autonomous Vehicles
by Samuel Oludare Bamgbose, Xiangfang Li and Lijun Qian

PAPR and BER Performances of OFDM System with Novel Tone Reservation Technique Over Frequency Non-Selective Fading Channel
by Moftah Ali, Raveendra K. Rao and Vijay Parsa

Critical Embedded Systems Development Using Formal Methods and Statistical Reliability Metrics
by Jonathan Lockhart, Carla Purdy and Philip Wilsey

An Improved Cross-Connection Abatement Algorithm with RSSI Using In-Band Magnetic Field Control in Densely Located LC Wireless Charger Environments
by Nam Yoon Kim, Jinsung Cho and and Chang-Woo Kim

Observing and Forecasting the Trajectory of the Thrown Body with use of Genetic Programming
by Konstantin Mironov, Ruslan Gayanov and Dmiriy Kurennov

A New Model of Supplier's Selection for Sustainable Supply Chain Management
by Oussama El Mariouli and Abdellah Abouabdellah

3D Reconstruction of Monuments from Drone Photographs Based on The Spatial Reconstruction of The Photogrammetric Method
by Andras Molnar

Performance Investigation of Semiconductor Devices using Commutation-speed based methodology for the application of Boost Power Factor Correction
by Barkha Parkash, Ajay Poonjal Pai, Wei Tian and Ralph Kennel

Strategies of the Level-By-Level Approach to the Minimal Route
by Nikolay Starostin and Konstantin Mironov

Enabling the Edge – A method for dynamic virtualizable connections for 5G deployments
by James Robert Pogge and Stephen lucien Scott

Over-The-Air Testing of Automotive Antennas and Wireless Links in The Installed State on The Basis of LTE Downlink Communication Parameters
by Philipp Berlt, Lisa Jäger, Andreas Schwind, Frank Wollenschläger, Christian Bornkessel and Matthias A. Hein

Talk Show's Business Intelligence on Television by Using Social Media Data in Indonesia
by Eris Riso and Abba Suganda Girsang

Emergence of Fun Emotion in Computer Games -An experimental study on fun elements of Hanafuda-
by Yuki Takaoka, Takashi Kawakami and Ryosuke Ooe

An Innovative EPC Architecture based on Not Only Stack for beyond 5G Mobile Networks
by Binwei Wu, Lu Ge, Jie Zeng, Xiangyun Zheng, Youxi Tang and Xin Su

Managing and Optimizing Quality of Service in 5G Environments Across the Complete SLA Lifecycle
by Evgenia Kapassa, Marios Touloupou, Panagiotis Stavrianos, Georgios Xylouris and Dimosthenis Kyriazis

Recent Advances in Intelligent-Based Structural Health Monitoring of Civil Structures
by Satyam Paul and Raheleh Jafari

Zebrafish Larvae Classification based on Decision Tree Model: A Comparative Analysis
by Bayan AlSaaidah, Waleed Al-Nuaimy, Mohammed Rasoul Al-Hadidi and Iain Young

A Practical Approach for Extending DSMLs by Composing their Metamodels
by Anas Abouzahra, Ayoub Sabraoui and Karim Afdel

Modeling an Energy Consumption System with Partial-Value Data Associations
by Nong Ye, Ting Yan Fok and Oswald Chong

Effects of The Low-Resolution 3D Video Clip on Cerebrum Blood Flow Dynamics
by Toru Tanimura, Hiroki Takada, Akihiro Sugiura, Fumiya Kinoshita and Masumi Takada

Using Fuzzy PD Controllers for Soft Motions in a Car-like Robot
by Paolo Mercorelli

Two Degree-of-Freedom Vibration Control of a 3D, 2 Link Flexible Manipulator
by Waweru Njeri, Minoru Sasaki and Kojiro Matsushita

Effect of Various Parameters for Temperature Distribution in Human Body: An Analytic Approach
by Kabita Luitel, Dil Bahadur Gurung and Kedar Nath Uprety

Study on CD ROADM Contention Blocking
by Guangzhi Li, Kerong Yan, Li Huang, Bin Xia, Fanhua Kong and Yang Li

Real Time Eye Tracking and Detection- A Driving Assistance System
by Sherif Said, Samer AlKork, Taha Beyrouthy, Murtaza Hassan, OE Abdellatif and M Fayek Abdraboo

CNN-based Automatic Coating Inspection System
by Lili Liu, Estee Tan, Zhi Qiang Cai, Xi Jiang Yin and Yongda Zhen

Difference in Speech Analysis Results by Coding
by Yasuhiro Omiya, Naoki Hagiwara, Takeshi Takano, Shuji Shinohara, Mitsuteru Nakamura, Masakazu Higuchi, Shunji Mitsuyoshi, Hiroyuki Toda and Shinichi Tokuno

Omni-directional Dual-Band Patch Antenna for the LMDS and WiGig Wireless Applications
by Mourad S. Ibrahim

Building an Online Interactive 3D Virtual World for AquaFlux and Epsilon
by Omar Al Hashimi and Perry Xiao

PID-Type FLC Controller Design and Tuning for Sensorless Speed Control of DC Motor
by Abdullah Y. Al-Maliki and Kamran Iqbal

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 201-210 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Experimental Investigation of Human Gait Recognition Database using Wearable Sensors

Sherif Said[*,1], Samer Al-Kork[1], Vishnu Nair[1], Itta Gowthami[3], Taha Beyrouthy[1], Xavier Savatier[3], M Fayek Abdrabbo[2]

[1]*American University of Middle East, Egaila, Kuwait*

[2]*Shoubra Faculty of Engineering, Benha University Shoubra, Cairo, Egypt*

[3]*IRSEEM, ESIGELEC Saint-Etienne du Rouvray, France*

ARTICLE INFO

Article history:
Received: 01 July, 2018
Accepted: 26 July, 2018
Online: 30 July, 2018

Keywords:
Gait Recognition
Biometrics
Wearable sensors
Accelerometer
Shimmer

ABSTRACT

In this research human gait database is collected using different possible methods such as Wearable sensors, Smartphone and Cameras. For a gait recognition accelerometer data from wearable shimmer modules and smartphone are used. Data from different sensors location is compared to know which sensor location have better recognition rate. Different walking scenarios like slow, normal and fast walk were investigated. Wearable sensors and smartphone data are compared to know whether mobile phones can be used for gait recognition or not. Also effects of age, height, weight on gait recognition are also studied. The obtained results of gait biometric matrices like Genuine Recognition Rate (GRR), Total Recognition Rate (TRR) and Equal Error Rate (EER) showed better results. EER in different walking scenarios ranged from 0.17% to 2.27% for the five wearable sensors at different locations, whereas EER results of smartphone data ranged from 1.23% to 4.07%. For sensors located at leg, pocket and hand the average GRR value falls with increase in age group, while for sensors located at upper pocket and bag, the GRR value doesn't follow any trend. Moreover GRR results on all sensors show no significance regarding height or weight variations.

1. Introduction

This paper is an extension of work originally presented in BioSMART, the 2nd International Conference on Bio-engineering for Smart Technologies" titled 'Biometric Database for Human Gait Recognition using Wearable Sensors and a Smartphone' [1]. Biometrics identifiers are typical, quantifiable characters that can be used to identify and label individuals. The identifiers can be either physiological or behavioural. Physiological characteristics include, but are not limited to fingerprint, iris recognition, face recognition, retina, and palm print etc., everything related to shape of the body. Behavioural characteristics include, but are not limited to gait, voice, typing rhythm etc. They are related to pattern of behaviour of a person. Biometric identifiers are unique for each person and they can be used as a reliable means to verify identity or as a means of authentication. But collection of biometric identifiers might raise concerns about privacy and questions about how secure the collected data is. Extensive use of biometric systems has been done in different fields such as forensics for criminal identification, electronic gadgets access, human activity recognition, health status [2]. Although extensive research has been going on in the field of biometric identification and authentication for the last decade, all this has been limited to the topics of face recognition, iris recognition, voice recognition, fingerprint recognition etc. Identification of individual using their gait is an idea which is least explored and not put into practise extensively [3]. Gait recognition is defined as "automatic identification of an individual based on the style of walking" [3]. Gait is the manner in which a person walks and it is more distinctive than we realise and so a person can be identified using his walking pattern from a distance [4]. Lot of research and thought has been put into human gait recognition using floor sensors (FS), machine vision (MV) and wearable sensors (WS) [5]. Most of previous works in gait recognition were based on machine vision techniques, i.e. analysing video or a sequence of images to collect patterns. Both MV and FS based techniques have their disadvantages. MV based techniques have many interfering variables [6] and FS based technique have costly floor sensors. Recently WS using accelerometers were used for gait recognition

[*]Sherif Said, American University of the Middle East, +965 2225 1400 Ext. 2168, sherif.said@aum.edu.kw

and nowadays every smartphone has an accelerometer giving a new direction to the gait recognition researches. Our model incorporates different widely used emerging technologies in human gait recognition. Wearable sensors, and Smartphone accelerometer two such technologies. In our experiment thought has been given to various walking scenarios and other variables such as age, weight, and height. The data acquired from the various sensors are then compared with testing data set. Different comparison methods give the matching percentage and best suited gait data. Figure 1 shows the outline of the data process that takes place in our work.

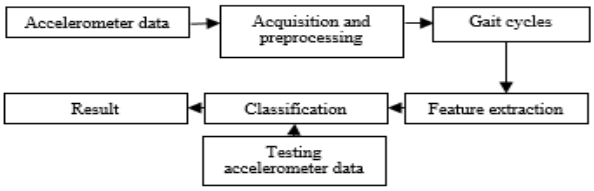


Figure 1. Overall process of gait cycle signal processing

2. Methodology

2.1. Database Collection

In a hall of 10x25 meters (indoor experiment), human gait database was collected using five wearable shimmer [7] sensor modules (integrated 3 axis MEMS accelerometer and gyroscope), and one smartphone Samsung Galaxy Note (with inbuilt NT70000 K3DH acceleration and K3G gyroscope sensors) [8], [9]. Data collected from these three techniques are saved with details like participant name, participant ID and type of walk in .dat, .csv, and .jpg formats respectively for further processing. None of the subjects has any known gait abnormalities. Figure 2 shows the overall setup for gait data collection. In this section we will talk in details about our experimental setup and collected methodologies.

Figure 2. Experimental lab used in our study

2.2. Wearable Accelerometer and Gyroscope Sensors

From a group of 50 people comprising of 37 males and 13 females of varying age range from 14 to 52 years' accelerometer and gyroscope data regarding gait is collected. The average age, height and weight of subjects is 26.6 years, 173.8cm and 71.2 Kg respectively. The general information regarding each subject like name, age, height, weight, and footwear are collected and kept securely for further data analysis. A walking protocol was developed and each subject were follow that protocols, which includes walking with different speeds like slow, normal and fast walk throughout the experiment. i.e. subjects should wait for 3 seconds before they start walking, then walk for a distance of four meters, then wait another 3 seconds and then walk the same distance to and fro for a duration of 45 seconds. The above same procedure was repeated for all other three types of walk.

Five wireless sensors modules (Shimmer 2r) are attached to different locations on human body like on L/R hand, L/R leg, L/R pant pocket, L/R shirt pocket and hand bag (Figure 3). The sensor locations for this study are chosen based on where a normal person carries his phone (like in pant pocket, hand, bag, upper pocket).

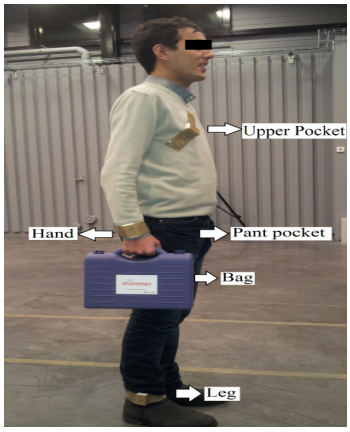


Figure 3. Shimmer sensor locations

Shimmer sensor module is a small wireless sensor platform that can record and transmit physiological and kinematic data in real-time (Figure 4). Each shimmer wireless sensor has on-board microcontroller (MSP430), wireless communication via Bluetooth or 802.15.4 low power radio and local storage to micro SD card [10], [11]. The unit also has integrated 3 axis MEMS accelerometer (Free scale MMA7361) and gyroscope for motion sensing, activity monitoring and inertia measurement application. All five Shimmer sensor modules are calibrated using 'Shimmer 9DOF Calibration application'.

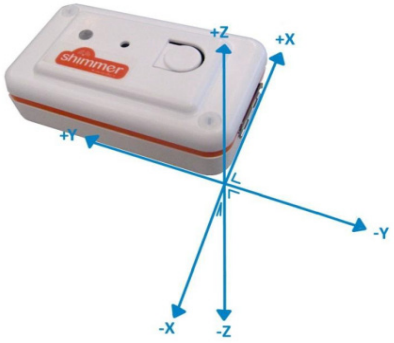


Figure 4. Shimmer 2R wearable sensor

Data acquisition software named 'Multi-Shimmer Sync' developed by shimmer research group was used to synchronize all five shimmer modules and to transmit streamed data to PC through Bluetooth. Multi-Shimmer Sync is an application which allows for the configuration and synchronized data capture from multiple Shimmers (Figure 5) [12], [13]. All five shimmer sensor modules are configured with ±1.5g acceleration range, this range is chosen because most of the smartphone's inbuilt accelerometers have same range. Sampling rate is set to 51.2 Hz

to avoid any data transmission loss. Total data collection time for each subject is approximately 4.5 minutes. Collected data are finally stored in PC in .dat format for further processing in MATLAB.

2.3. Smartphone accelerometer and Gyroscope

From a group of 23 subjects comprising of 16 males and 7 females of a varying age group from 21 to 39 years. The average age, height and weight of collected subjects is 27 years, 172.2cm and 72.08 Kg respectively. Smartphone Samsung galaxy note, with inbuilt NT70000 K3DH acceleration [14] and K3G gyroscope sensors are used to capture data [15]. Android application named 'Sensor pro list' is used to capture sensor data. Captured sensor data is then transferred to PC through Bluetooth. [16], [17]. Each subject is asked to hold the phone in hand like how they normally carry it [18]. Each subject is also asked to select the log on and log off (i.e. to start and stop) of mobile android application and walk in similar walk protocol as explained earlier [19], [20]. Table 1 shows the comparison study of database collected using two techniques.

3. Processing of Accelerometer Data

Raw data from all shimmer sensors modules is saved in DAT file format. Each subject file contains information of the five sensors. Each sensor data consists of time stamp, accelerometer

(in x, y, and z axis) and gyroscope (in x, y, and z axis). similarly, smartphone data is a saved in CSV file format. Each subject smartphone sensor data also consists of time stamp, accelerometer (in x, y, and z axis) and gyroscope (in x, y and z axis).

3.1. Data Preprocessing

3.1.1 Data reading

The data reading procedure is divided into three steps. In first step, each participant data such as name, ID, height, weight, gender, age, sensor location is read one by one which are stored in excel file. In second step, sensor data files of each participant are read from assigned folders and subfolders for creation of gait data base features. Finally, all data files are exported to MatLab and headers of each file are read to identify each subject recorded

Figure 5. Sample gait data collection with Multi shimmer sync software

Table 1. Comparison of the two techniques

	Wearable Shimmer sensor	*Smartphone*
Sensor specification	Five Shimmer 2r's with inbuilt 3-axis MEMS Freescale MMA 7361 accelerometer	One Samsung Galaxy Note (inbuilt K3DH acceleration & K3G gyroscope sensors)
Computer interface	Bluetooth	Bluetooth
Software used	Multi Shimmer Sync	Sensor Pro list
Range of sensor	+/-1.5 g	+/-1.5 g
Sensor sampling frequency	51.5 Hz	
Location of sensor	Leg, Hand wrist, pant pocket, Shirt pocket and bag (left and right side)	Hand
Number of participants	50(37 male, 13 female)	23 (16 male, 7 female)
Range of age	14 to 52 years	21 to 39 years
Mean of age	26.6 years	27
Range of height	150-193 cm	154-189 cm
Mean height	173.8 cm	172.2 cm
Range of weight	50 to 107 kg	50 to 105 kg
Mean weight	71.2 kg	72.08 kg

sensor data belong to which sensor location (like leg data\hand data\pocket data etc.) and to which data such as accelerometer, gyroscope and time stamp values. Since each sensor at different body location has different processing techniques such as data recorded from leg is entirely different from data recorded from pocket; therefore, data each particular sensor is identified and processed with techniques explained below. Processing methods also depends on the sensors used for data collection, like sampling frequency of sensor, location of sensor. Effective preprocessing techniques improve recognition rate. Figure 6 shows a flow chart of data processing of testing and training data.

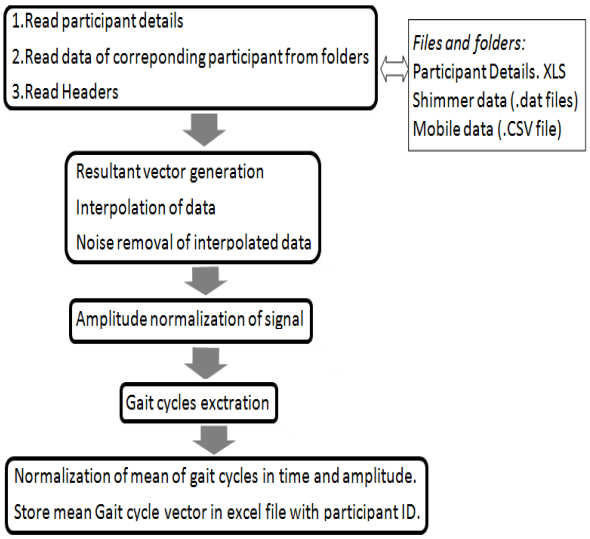


Figure 6. Flow chart of data processing of testing and training data

3.1.2 *Data processing*

Resultant vector: The output from various accelerometers data will vary depending on how the shimmer sensor modules are oriented and also all three different axes have varying moments. To overcome this problem, the resultant vector (xyz) of the accelerometer in X, Y, and Z axis output is calculated using Euclidean norm as given in (1). The first three plots of Figure 7 shows X, Y, Z plots. The last plot in Figure 7 shows the resultant vector of all axes.

$$xyz = \sqrt{x^2 + y^2 + z^2} \quad (1)$$

Interpolation: Raw accelerometer data has irregular periodic intervals therefore interpolation is needed to have data samples at regular intervals. Interpolation [21] is a method of constructing new data points within the range of a discrete set of known data points. Many interpolation methods like linear, polynomial, spline interpolation methods can be used. In our work, spline interpolation of period 10 m sec is used. Spline interpolation [22] uses low-degree polynomials in each of the intervals, and chooses the polynomial pieces such that they fit smoothly together [23].

Noise removal: Weighted moving average (WMA) method is applied to our interpolated data to remove unwanted noise from the signal. This method is fast and easy to implement. In WMA method, the nearest neighbors are more important than those more away, while in other methods all the neighbors have equal weight. The formula for WMA with a sliding window of size 5 is given in (2).

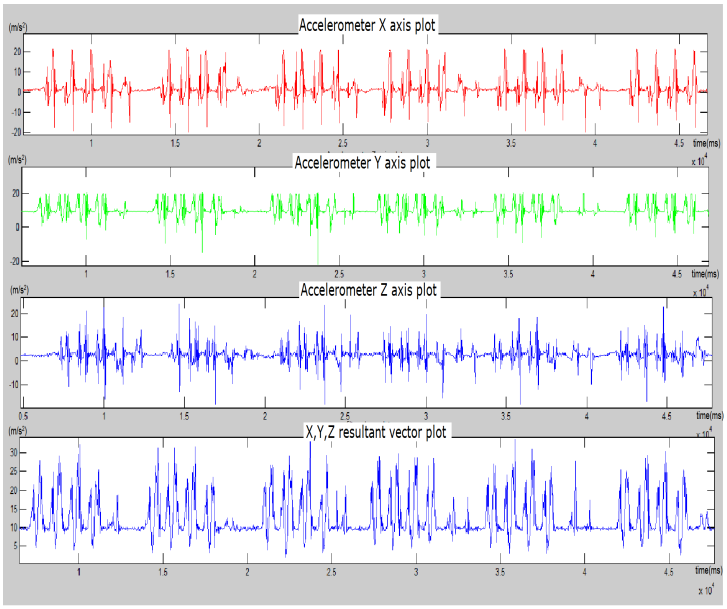


Figure 7. Unprocessed raw gait signals

$$wma = \frac{(x_{t-2}*1)+(x_{t-1}*2)+(x_t*3)+(x_{t+1}*2)+(x_{t+2}*1)}{5} \quad (2)$$

where: x_t is the acceleration value at position t

Amplitude Normalization: For easy computations, raw accelerometer data is normalized from 0 to 1 as shown in Figure 8 by applying (3)

$$\text{newsi} = \frac{\text{xyzif}-\text{mi}}{\text{ma}-\text{mi}} \quad (3)$$

Where: newsi - amplitude normalized signal

xyzif - interpolated and filtered signal of accelerometer resultant vector signal

mi - minimum range of sensor

ma - maximum range of sensor

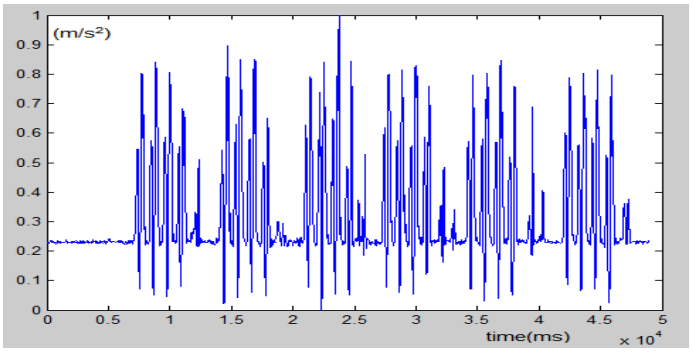


Figure 8. Amplitude normalization of whole gait signal

3.1.3 Gait cycles extraction and Time Normalization

Cycle Detection: A gait cycle will comprise of two footsteps and it is detected by finding the minimum points in a given cycle.

The end of the preceding one will mark the start of each cycle up to final cycle of the gait signal. Fake minimum points are eliminated by calculating mean cycle time and standard deviation. If a gait cycle has a cycle length which falls outside of the mean time ± standard deviation is considered to be as fake cycles and will be eliminated. Since all sensors collect gait data whose minimum points should mark the start of another gait cycle, the gait detection algorithm will work for all the sensors used. In Figure 9, the first plot shows extracted gait cycles from an amplitude normalized signal. Cycle length as well as mean length of all gait cycles are estimated and saved as feature vectors. There are chances that the length of each gait cycle might vary from cycle to cycle; therefore, normalization of signal in time is needed. Gait cycles are normalized to 1 second duration. Here 1 second is chosen as a random standard value. Figure 9 shows difference between normalized and un-normalized gait cycles.

In some cases, after time normalization also fake cycles are observed, so again these fake cycles are eliminated by calculating trimmed mean cycle (TM cycle). TM cycle is calculated by calculating mean and standard deviation (SD) of all gait cycles and if point lies beyond ±SD of mean cycle then that particular cycles eliminated. Figure 10 shows a normalized trimmed gait cycle.

3.1.4 *Normalization of Mean gait*

After elimination of fake cycles, Mean cycle, Median cycle and Trimmed mean cycle are calculated to represent a gait cycle feature vectors. Mean cycle refers to the time domain waveform at which each time sample represents the mean cycle value over a fixed time window of the gait cycle length. Median cycle refers to the time domain waveform at which each time sample represents the median cycle value over a fixed time window of the gait cycle length. Mean, Median and TM gait cycle which is normalized in time (for 100 samples) and in amplitude (0 to 1 range) is then stored in an array of 100 values are stored in Microsoft excel.

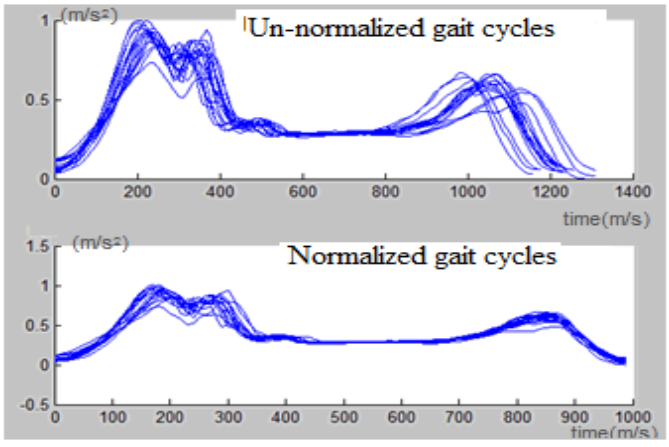


Figure 9. Un-Normalized and Normalized gait cycles

Since each sensor at different body location has different processing techniques such as data recorded from leg is entirely different from data recorded from pocket; therefore, each sensor is identified and processed with techniques explained below. Processing of different sensors varies in cutoff and in elimination of fake cycles techniques.

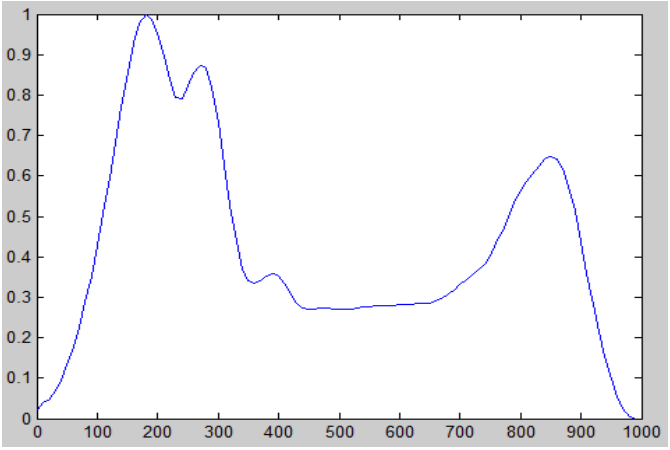


Figure 10. Normalized Trimmed Mean Gait Cycle

3.2. *Elimination of fake cycles*

Before calculating average gait cycle fake cycles can be eliminated by using following techniques:

1. By calculating mean cycle length(MCL) and standard deviation (SD) of all identified cycles. Cycle length falling beyond MCL ± SD are eliminated.
2. By calculating trimmed mean(TM) cycle of all cycles. TM cycle is obtained by calculating mean of cycle points which lie in between Mean cycle point ± SD of cycle point and cycle points which lie beyond range are eliminated.
3. By matching maximum point of all gait cycles i.e. by calculating mean gait cycle and searching maximum point of each cycle around SD of mean gait cycle maximum point.

3.3. *Data analysis for obtaining results*

This section explains how data has been analysed for obtaining results.

3.3.1 *Analyzing methods*

This sub section explains the calculation of gait biometric system performance metrics like FAR,FRR,EER,TRR and FRR [24]. The performance metrics are calculated by creating Distance score tables and explained below.

Distance score table calculation: Obtained average gait cycles for all subjects for different walks(SW, NW, FW) at different sensor location for both smartphone data and wearable data are compared against each other. Distance metrics like Manhatten and Euclidean methods are used and explained below:

Manhatten distance: This is also known as absolute distance and the formula is shown in (4). Manhattan distance between two points is the sum of the absolute difference in their Cartesian coordinates. In addition, this distance metric is the computationally least expensive one.

$$dm = \sum |(x1 - x2)| \tag{4}$$

Hamming distance: Hamming distance is utilized to detect and correct errors in digital communication. Hamming distance between two data are said to be the least number of changes that could make both the data same. For example, the hamming

distance between 'name" and 'meme" is 2 and between 337895 and 235817 is 4.

Euclidean distance: This is a slight modification of the Manhattan distance, see (5). Instead of taking the sum of the absolute differences we now take the square root of the sum of all differences squared.

$$dh = \sqrt{\sum(x1 - x2)^2} \quad (5)$$

From distance metric methods a score tables are generated. Valid subject has a less distance score as compared to not valid subjects. Distance score table is classified into accepted and rejected matches based on classifier cutoff. Classifier cutoff is choosen from percentage of maximum score value from score table. Accepted matches again have two cases like Genuine Accepted Match (GAM) and Fraud Accepted Match (FAM). GAM is accepted match of correct subject i.e. accepted match is of correct subject and is recognized by classifier. FAM is counted when a false match is accepted and recognized as correct match. Similarly rejected matches have two cases like Genuine Rejected Match (GRM) and Fraud Rejected Match (FRM). GRM is genuine match is supposed to be accepted but is not recognized. FRM is counted when false match is accepted and recognized as an incorrect match. FRM is fraudulent match is supposed to be accepted but is recognized as incorrect.

Biometric matrices like False acceptance rate (FAR), false reject rate (FRR), Total recognition rate (TRR), Genuine recognition rate (GRR) and Equal error rate (ERR) are calculated from generated score table and are explained as below.

False acceptance rate (FAR) is a measure of how precisely biometric data can be compared and recognized (6). It represents the chance that the comparison will accept a wrong input as an affirmative match. The input is not supposed to match with the template data but invariably the system considers this match to be correct. FAR is calculated by testing known biometric templates against a huge collection of data.

$$FAR(\%) = \frac{number\ of\ fraud\ accepted\ samples(FAM)}{total\ number\ of\ samples} * 100 \quad (6)$$

False rejection rate (FRR) is a measure of the chance that the system will wrongly reject a genuine input as a match that doesn't fit (7).

$$FRR(\%) = \frac{number\ of\ genuine\ rejected\ samples(GRM)}{total\ number\ of\ samples} * 100 \quad (7)$$

EER is the value at which FRR and FAR are equal and is obtained by plotting graph between FAR and FRR. TRR and GRR are calculated using (8) and (9). In TRR number of recognized samples is sum of genuine accepted and fraud rejected samples but GRR is calculated only for genuine accepted samples.

$$TRR(\%) = \frac{number\ of\ recognised\ samples(GAM+FRM)}{total\ number\ of\ samples(N)} * 100 \quad (8)$$

Where N is the total number of comparison samples.

$$GRR(\%) = \frac{number\ of\ genuine\ recognised\ samples(GAM)}{total\ number\ of\ samples} * 100 \quad (9)$$

For data collected from wearable sensor for 50 subjects, two files for each of three different walking scenarios (SW, NW and FW) we have total 300 (50*3*2) gait features for one sensor location. So total of 100 gait features of each scenario (e.g. for normal walk sensor at leg have 100 templates) are compared against each other and with other scenario i.e. comparison with walks like slow-slow(S-S), normal-normal (N-N), fast-fast (F-F), slow-normal(S-N), slow –fast(S-F), normal-fast (N-F). In comparison of same sensor location and same walk there are five cases, there are case 0: don't count (DC), case 1: GAM, case 2: GRM, case 3: FAM, case 4: FRM. DC case is when comparing two similar files because each subject has two files when we compare first file of a subject to the first file of subject it is obvious match. So this case is not counted for metric calculation. Also if there is match between first file of subject to second file this is counted in GAM. So cases 0, 2, 3 are not valid (error matches), FAR and FRR are calculated using FAM and GRM respectively. In comparison with other different walks like), slow-normal(S-N), slow –fast(S-F), normal-fast (N-F) case 0 (DC) will not be there.

Table 2 shows the FAR, FRR, TRR and GRR values with different classifier cutoffs. Consider FW-FW comparison with cutoff 40%, it has FAR of 2.92% which says that 2.92 subjects out of 100 subjects are falsely accepted, FRR of 0.07% means 0.07 subjects which are genuine are rejected, TRR of 97.01% says how efficiently a classifier can classify matches correctly whether it is genuine accepted match or fraud rejected match and GRR of 93% represent the ability of classifier to identify genuine accepted matches. From Table 2, as cutoff percentage increases GRR increases at the cost of increase in sum of errors (FAR+FRR). So classifier cutoff is chosen such a way by having tradeoff between GRR and errors.

Table 2. Pocket hamming method results

100 samples	Classifier Cut-off=10%				Classifier Cut-off=20%				Classifier Cut-off=30%				Classifier Cut-off=40%				Classifier Cut-off=50%			
Pocket	FAR	FRR	TRR%	GRR	FAR	FRR	TRR%	GRR	FAR	FRR	TRR%	GRR	FAR	FRR	TRR%	GRR	FAR	FRR	TRR%	GRR
S-S	0	0.67	99.3300	33.00	0.72	0.3900	98.8900	61.00	3.0800	0.2200	96.7000	78.00	6.2900	0.1200	93.5900	88.00	8.4500	0.0300	91.5200	97.00
s-n	2.3300	1.55	96.1200	22.50	7.30	1.0800	91.6200	46.00	12.9800	0.8500	86.1700	57.50	17.9900	0.6500	81.3600	67.50	20.4500	0.5300	79.0200	73.50
s-f	1.9700	1.77	96.2600	11.50	10.88	1.3200	87.80	34.00	23.6700	0.9100	75.4200	54.50	31.4400	0.6900	67.8700	65.50	34.3100	0.6000	65.0900	70.00
n-n	0.1800	0.22	99.6000	78.00	0.93	0.0800	98.99	92.00	1.8800	0.0500	98.0700	95.00	2.4800	0.0200	97.5000	98.00	3.0600	0.0200	96.9200	98.0000
n-f	5.5000	1.03	93.4700	48.50	12.77	0.6500	86.58	67.50	15.1800	0.5800	84.2400	71.00	17.9300	0.5100	81.5600	74.50	19.1300	0.4700	80.4000	76.50
f-f	0.0500	0.50	99.4500	50.00	0.63	0.1500	99.22	85.00	1.8800	0.0900	98.0300	91.00	2.9200	0.0700	97.0100	93.00	3.9100	0.0400	96.0500	96.00

4. Results

Figures 11 and 12 shows the recognition rate of all wearable sensors using Hamming method of classification. Figures 13 and 14 shows the recognition rate of all wearable sensors using Manhattan method of classification Same sensor to sensor walk (like S-S, N-N, F-F) comparison have highest recognition rate value than comparison with other walks (S-N, S-F, N-F). In Figures 11,12,13 and 14 shows that Manhattan classifier has better classification results than Hamming method. Sensor located at Pocket and Leg has highest recognition rate than other sensor location. Results in table shows the EER value of different sensors. The EER values ranges from 0.21 to 2.258. Same sensor to sensor walk EER values (for S-S, N-N, F-F) comparison have less error values than comparison with other walks (S-N, S-F, N-F). EER rate of F-F comparison is more than S-S and N-N because when a subject is asked to walk fast, walk becomes unstable. Also among all six scenarios S-F comparison has highest EER as there is template mismatching and comparison between stable (SW) and unstable walk (FW) gives higher EER.

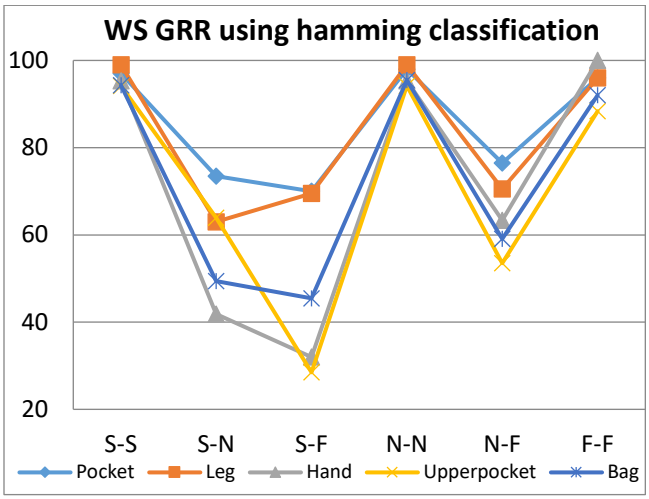


Figure 11. Wearable sensors GRR values using hamming method

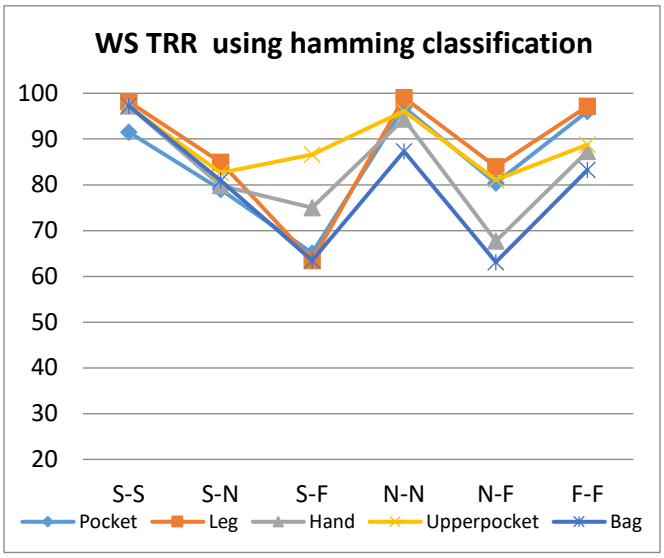


Figure 12. Wearable sensors TRR values using hamming method

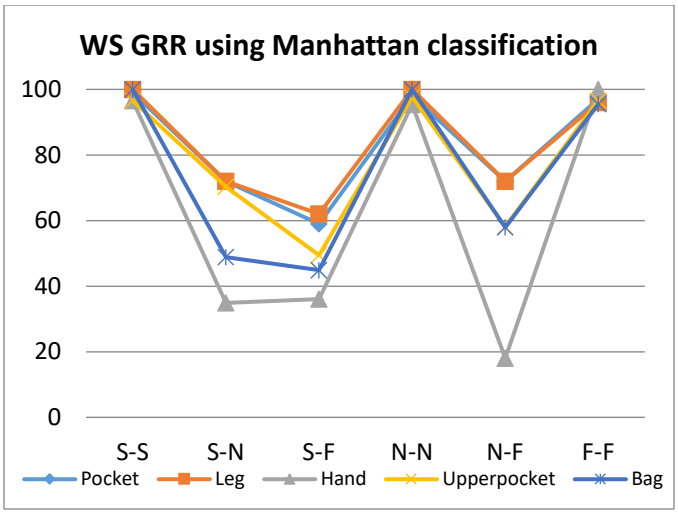


Figure 13. Wearable sensors GRR values using Manhattan method.

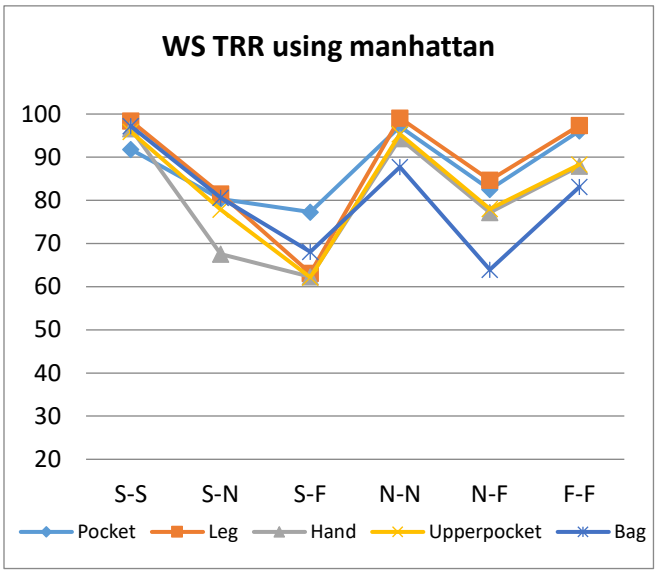


Figure 14. Wearable sensors TRR values using Manhattan method

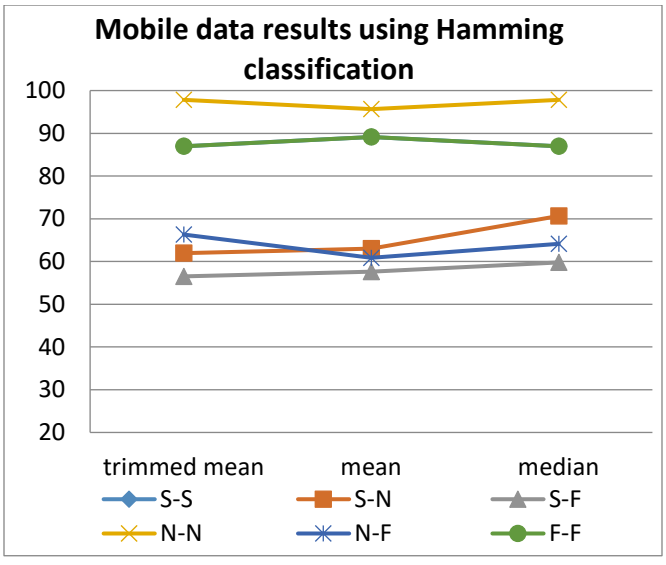


Figure 15. Mobile data results of Hamming method

Table 3. EER table of Hamming and Manhattan tables

	Hamming					Manhattan				
	Pocket	Leg	Hand	Upper Pocket	Bag	Pocket	Leg	Hand	Upper Pocket	Bag
S-S	0.44	0.2700	0.3245	0.3786	0.3874	0.4400	0.2700	0.3245	0.3786	0.3616
s-n	1.6200	1.6000	1.9605	1.3521	1.7691	1.6100	1.6100	2.2445	1.3386	1.7562
s-f	1.7900	1.9200	2.2039	2.0146	2.1436	1.7800	1.8900	2.2580	2.0687	2.1823
nn	0.2100	0.1700	0.4867	0.4056	0.8781	0.1900	0.1700	0.4867	0.4056	0.8781
n-f	1.3600	1.5000	2.0822	2.0416	2.0919	1.4600	1.5000	2.2715	2.0416	2.0919
f-f	0.2400	0.2600	0.5949	0.6760	0.8910	0.2300	0.2600	0.6084	0.6760	0.8910

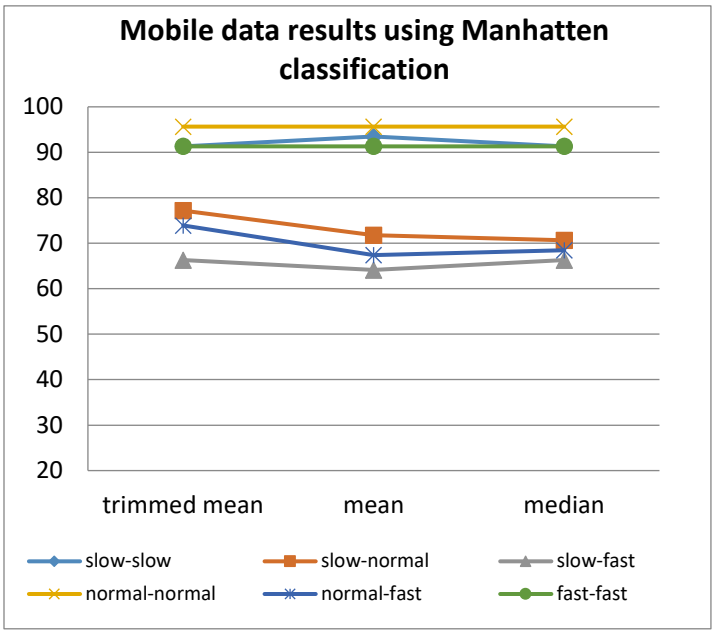


Figure 16. Mobile data results of Manhattan method

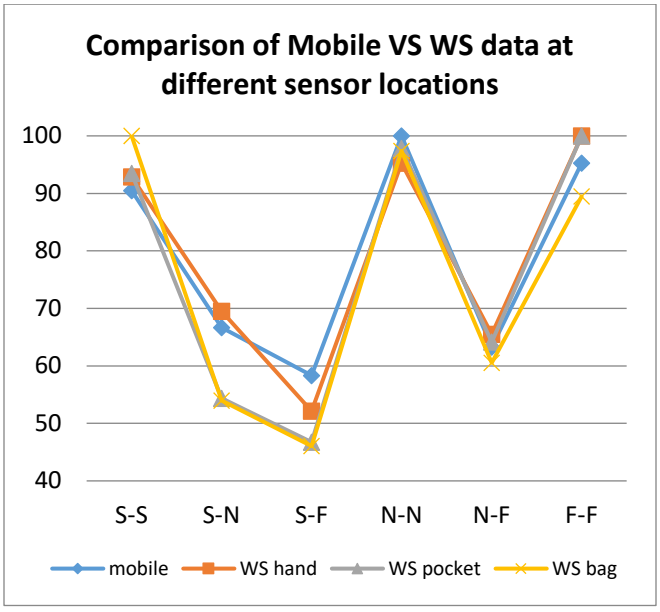


Figure 17. Comparison of mobile data results with wearable sensors location like hand, pocket and bag data

Figures 15,16 shows the results of Manhattan and Hamming methods for six different walk scenarios for which comparisons are done TM, mean and median cycles respectively. Table 3 shows the EER values, whose range is from 1.2287 to 4.0643. EER values of mobile data are more than wearable sensors data.

In real life people hold\place their phones in different location on body such as (pocket, hand, suit case etc...). Figure 17 shows the comparison results of GRR for 23 participants' data of mobile, wearable sensors at hand, pocket and bag. Even though mobile data have more or same wearable sensors recognition rates, mobile data has more EER values. Smartphone's have accelerometers that are good enough to detect human gait with slight modification on their frequency range and sensitivity of sensor. Although investigations are going on extensively in this field of research, an attempt to include comparison of different techniques and that to taking into account different walking scenarios was never seen before. This paper will give insight into how different walking posters will affect the gait recognition of different sensors.

Table 4. Mobile data EER rate

	Hamming			Manhattan		
	Trimmed mean	Mean	Median	Trimmed mean	Mean	Median
S-S	1.6068	1.7013	1.6068	1.6068	1.6068	1.6068
s-n	3.5917	3.6389	3.5444	3.5917	3.6389	3.5444
s-f	4.0643	4.0643	4.0170	4.0643	4.0643	4.0170
n-n	1.7486	1.7958	1.7958	1.6541	1.7486	1.7013
n-f	3.5917	3.6389	3.3554	3.5917	3.6389	3.3081
f-f	1.3233	1.2287	1.3233	1.2287	1.2287	1.2760

Human gait is affected by many factors, and changes in the normal gait pattern can be transient or permanent. The factors can be of various types:

- Extrinsic: Several extrinsic factors such as terrain, footwear, clothing, cargo(luggage).
- Intrinsic: Intrinsic factors are sex (male and female), weight, height, age, etc.
- Physical: Physical factors such is the weight, height, physique

- Psychological: Psychological factors are the type of personality, emotions
- Physiological: When talking about is physiological anthropometric characteristics
- Pathologic, pathological factors can be for example trauma, neurological diseases, musculoskeletal, psychiatric disorders
- Intrinsic factors such as age, height and weight are also analyzed in our work.

The entire sample was divided into three equal groups as shown in Figure 18 for comparing wearable sensors for different weight range. And it was found that the average TRR for range of 50 to 58 Kg is more than other two ranges.

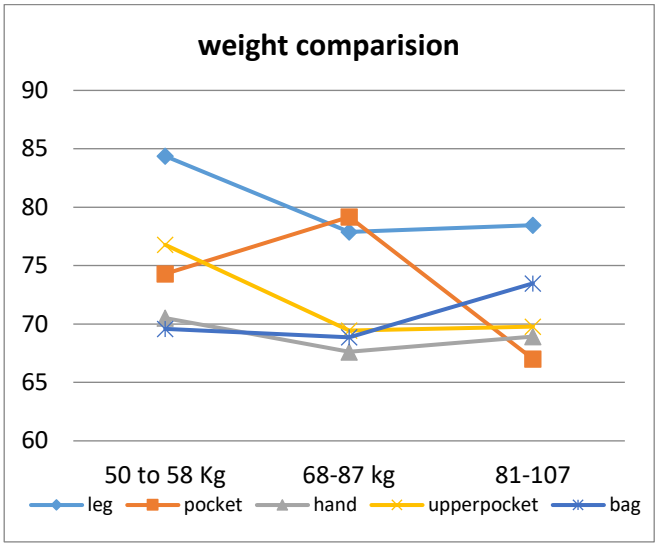


Figure 18. Average GRR (of six scenarios) value of wearable sensor data comparison for different weights range

For sensors located at leg, pocket and hand the average GRR value falls with increase in age group, while for sensors located at upper pocket and bag, the GRR value doesn't follow any trend. In general, average recognition rate decreases as age increases, which has been explained by Richard W. Bohannon [25] in his work. Many other physiological factors like brain stability, thoughts etc. changes as age increases (Figure 20). Weight and height factors are studied as shown in Figure 18 and Figure 19 respectively. Different walks had different energy of signal which can be used for activity recognition. Even though three walks have different energies, they have more or less same gait features in common.

5. Conclusion

Our future work needs to give emphasis to increasing the number of participants in the experiment to about 120-150 because the number of participants affect the calculated recognition rate by a great deal and helps in making the calculations accurate. Time elapsed and activity recognition are some other factors that will have to be pondered over. This research will provide us with an idea of how changes in subject's daily routine affect the rate of recognition. How a person walks can vary depending on different parameters such as footwear, type of clothing, with or without luggage, walking surface and terrain, human factors; thus the above mentioned issues needs to be looked into further. Our research needs to be further extended to include other types walking patterns including running, climbing stairs up/down, jumping and sitting. The smartphone data used in the study was exclusive to one smartphone and one data processing software. More work is to be done on this part. Data from different smartphones having a variety of built in accelerometers and using some other data processing software. Effect of Gait difference of same height subjects with different weight and effect of load are also to be studied further. Insight was also provided on how other co parameters affect the gait recognition of the sensors.

Figure 19. Average GRR (of six scenarios) value of wearable sensor data comparison for different heights range

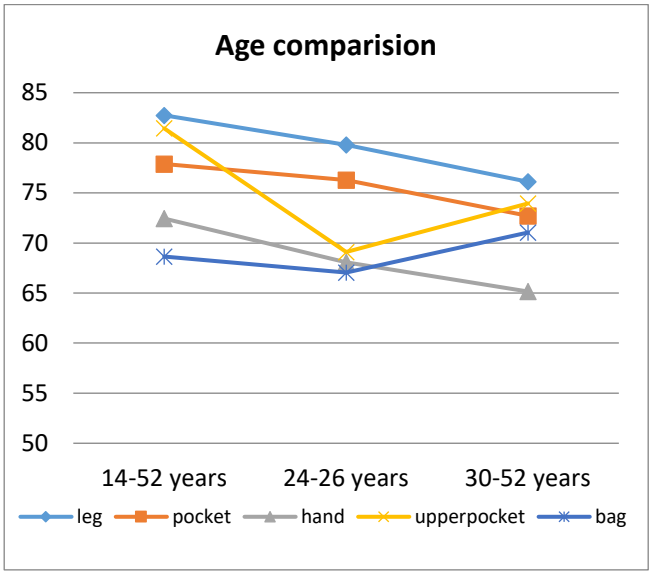


Figure 20. Average GRR (of six scenarios) value of wearable sensor data comparison for different weights range

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by Haute-Normandie Region as part of The Nomad Biometric Authentication (NOBA) project.

References

[1] S. K. A. Kork, I. Gowthami, X. Savatier, T. Beyrouthy, J. A. Korbane and S. Roshdi, "Biometric database for human gait recognition using wearable sensors and a smartphone," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4

[2] Qian Cheng, Yanen Li_, Joshua Juen, "Classifier of Gait Pattern for Detecting Abnormal Health Status", SenSys'12Toronto, ACM 978-1-4503-1169-4, Canada 2012.

[3] D. Hodgins, "The Importance of Measuring Human Gait," European Medical Device Technology, 2008.. F Sayegh, F Fadhli, F Karam, M BoAbbas, F Mahmeed, JA Korbane, S AlKork, T Beyrouthy "A wearable rehabilitation device for paralysis," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp.1-4.

[4] Mark Ruane Dawson, "Gait Recognition",(technical report) Department of Computing Imperial College of Science, Technology & Medicine London 2002.

[5] Q. Meng b, B. Li a, H. Holstein, "Recognition of human periodic movements from unstructured information using a motion-based frequency domain approach," Volume 24, Issue 8, Pages 795–809, Department of Computer Science, University of Wales, August 2006.

[6] T. Massart, C. Meuter and V. B. Laurent, "On the complexity of partial order trace model checking, Inform. Process," 2008.

[7] M. Davis and H. Putnam, "A computing procedure for quantification theory.," in Journal of ACM, 1960

[8] E. Abraham, S. Alkork, F. Amirouche, E. Hande, S. P. Rai and S. Biafora, "Experimental Investigation of Posterior Elbow Dislocation in a Primate Model," 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, 2007, pp. 5099-5102

[9] 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 2009 Summer Bioengineering Conference, Parts A and B ():625-626

[10] S. Said, S. AlKork, T. Beyrouthy and M. F. Abdrabbo, "Wearable bio-sensors bracelet for driver as health emergency detection," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSmart), Paris, 2017, pp. 1-4.

[11] F Sayegh, F Fadhli, F Karam, M BoAbbas, F Mahmeed, JA Korbane, S AlKork, T Beyrouthy "A wearable rehabilitation device for paralysis," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp.1-4.

[12] Ho, N. S. K., Tong, K. Y., Hu, X. L., Fung, K. L., Wei, X. J., Rong, W., & Susanto, E. A. (2011, June). An EMG-driven exoskeleton hand robotic training device on chronic stroke subjects: task training system for stroke rehabilitation. In 2011 IEEE

[13] Taha Beyrouthy, Samer Al. Kork, J. Korbane, M. Abouelela, "EEG Mind Controlled Smart Prosthetic Arm – A Comprehensive Study", Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 3, pp. 891-899 (2017).

[14] "http://www.shimmer-research.com.," [Online].

[15] Sung-Jung Cho, Eunseok Choi, Won-Chul Bang, Jing Yang, "Two-stage Recognition of Raw Acceleration Signals for 3-D Gesture-Understanding Cell Phones," Tenth International Workshop on Frontiers in Handwriting Recognition 2006.

[16] M. O. Derawi, "Accelerometer-Based Gait Analysis A survey," in NISK conference., 2010.

[17] Mohammad Omar Derawi, "Accelerometer-Based Gait Analysis A survey," pp 33-44, NISK-2010 conference.

[18] Tobias Stockinger, " Implicit Authentication on Mobile Devices," Ubiquitous Computing, Media Informatics Advanced Seminar LMU, Summer Term 2011(Tech Report), University of Munich, Germany 2011.

[19] Jani Mäntyjärvi, Mikko Lindholm, "Identifying Users of Portable Devices from Gait Pattern with Accelerometers," pp. ii/973 - ii/976 Vol. 2, IEEE 2005.

[20] Yuzuru Itoh, Manabu Miyamoto, "Convenient use of Accelerometer for Gait Analysis," pp 21-31, Bulletin of the Osaka Medical College, 2008

[21] M. Lindholm and J. Mäntyjärvi, "Identifying Users of Portable Devices from Gait Pattern with Accelerometers," IEEE, vol. Vol 2, pp. 973-976, 2005.

[22] S.-J. Cho, E. Choi, W.-C. Bang and J. Yang, "Two-stage Recognition of Raw Acceleration Signals for 3-D Gesture-Understanding Cell Phones," in Tenth International Workshop on Frontiers in Handwriting Recognition, 2006.

[23] Jordan Frank, Shie Mannor, "Activity and Gait Recognition with Time-Delay Embeddings," Twenty-Fourth AAAI Conference on Artificial Intelligence 2010.

[24] Martin Hynes, Han Wang, "Off-the-Shelf Mobile Handset Environments for Deploying Accelerometer based Gait and Activity Analysis Algorithms," Engineering in Medicine and Biology Society, pp. 5187 – 5190, IEEE 2009.

[25] R. W. Bohannon, "Comfortable and maximum walking speed of adults aged 20—79 years: reference values and determinants," 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 01-07 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Automation System for Regulation Optimization in Power Transformer Design

Tatjana Šimović[*,1], Mislav Gazdović[2]

[1]*Končar Power Transformers Ltd., A joint venture of Siemens and Končar /Zagreb 10000, Croatia*

[2]*Montelektro d.o.o./Kastav 51215, Croatia*

ARTICLE INFO

Article history:
Received: 29 November, 2018
Accepted: 17 December, 2018
Online: 11 January, 2019

Keywords:
Power transformer regulation
Tap changer selection
Tap changer database

ABSTRACT

Large power transformers generally include a customer request for a technically appropriate regulation unit. The selection process of the regulation unit consists of defining the required input data, performing mathematical calculations necessary to find the technical limit values that the regulation unit has to satisfy, and finally optimizing and selecting the appropriate regulation unit. The number of possible permutations consists of thousands of different combinations, depending on the type of regulation and other technical limitations. The automation system presented in this paper significantly reduces the time required to obtain the optimal regulation unit solution in both offer and order stages of the project, providing significant overall productivity increase. This paper presents an example of managing a part of a complex system such as power transformer design using a software solution. The process of finding the optimal regulation unit „manually" can take up to several hours. Implementing the developed algorithm and introducing Tap Changer Selection application, the required time is reduced to several minutes. This represents significant time savings and reduced possibility of errors, thus improving the power transformer design process.

1. Introduction

This paper is an extension of work originally presented in the conference Proceedings of the 41st International Convention for Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018) [1].

Large power systems such as electrical grids of entire cities, countries or continents are systems which must be continuously regulated in order to keep the voltage supply constant, thereby ensuring all the electrical devices and equipment to work in the expected voltage range. In such large systems, disturbances in the load and/or supply are inevitable – i.e. under greater load, the current increases, and consequently the voltage drops along the transmission lines which connect the power supply to the power consumption. It is important to emphasize that there are two main types of voltage deviations. The first is caused by a change of load on the lower voltage side of the power transformer. Since load through the power transformer has changed, the change of the position of the regulation unit is required in order to maintain the voltage at a constant level. The second voltage disorder is the case when voltage is changed on the higher voltage side of the power transformer as a consequence of a certain kind of network disturbances [2]. Handling such disturbances and maintaining a constant voltage supply is (among others) the responsibility of the power transformer as one of the most expensive parts of the power grid. For this purpose, the regulating power transformer is equipped with a regulating winding which is connected to the regulation unit - as shown in Figure 1 - usually On Load Tap Changer (OLTC).

The OLTC principle was patented in 1927 by Dr. Bernhard Jansen. Together with the industrial development and economic growth over the years, the overall power consumption took an upward trend, following the expansion of the electrical grid. The development of OLTCs was accelerated over the years due to the steady increase of the transmission voltage and power. The voltage regulation basic principle is adding or subtracting turns from either the primary or the secondary winding. The OLTC alters the power transformer turns ratio in a number of predefined steps and in that way changes the secondary (or primary) side voltage [3].

As opposed to the OLTC, the DETC (De Energized Tap Changer) must not be operated while the transformer is energized.

[*]Corresponding Author: Tatjana Šimović, Zagreb 10000, Croatia, +385989889304, Email: tatjana.simovic@siemens.com

This means that for the DETC to operate, the transformer must be disconnected from the network for a short period of time. Due to the mentioned limitations, the DETC is used only when the regulation capabilities are not a primary concern. This is why an OLTC regulation unit is more and more becoming a standard customer request for power transformer regulation.

Figure 1: Power transformer with a regulation unit

Generally, many different regulation units are technically appropriate and may be used for a particular transformer design. In order to minimize the overall costs while still managing to provide the required regulation capabilities, one of the transformer manufacturers' challenges is to optimize the regulation unit selection process. Since the regulation unit (i.e. tap changer) can make around 10 – 20% of the total material cost of the transformer, the overall effect of optimizing the regulation unit becomes significant.

Related work on this area consists of different power transformer design tools which (at most) consist of a regulation unit database, expecting the user to choose the appropriate regulation unit from the list (manually). A tool example presenting automated regulation unit selection functionality in the power transformer industry is ABB Compas tool [4]. Since this tool is limited only to ABB types of tap changers, the main motivation for this work was to provide a faster and more efficient way of selecting the optimal regulation unit also for manufacturers other than ABB, e.g. MR (Germany) and C.A.P.T. (Italy) for OLTC and DETC regulation units since this is a common customer request. The main advantage of this application becomes evident in the offer stage where quick and correct selection of the optimal regulation unit is important for the final transformer design cost estimation.

2. Transformer Regulation

The transformer is an electrical static device without moving parts, used to transform electrical power from one circuit to another without electrical connection and without changing frequency. Transformer switches the alternating current of the predetermined electrical voltage into alternating current of a higher or lower electrical voltage using the effect of mutual induction. The only moving part in a power transformer, the On Load Tap Changer, is one of the main contributors to the failure rates of high voltage power transformers [5]. Taking this into consideration, it makes sense to standardize and automate the regulation unit selection process, which was the main motivation for this work.

The introduction of OLTCs improved the operating efficiency of electrical systems considerably and this technique found acceptance worldwide. In general the percentage of transformers equipped with OLTCs is increasing with the increase of the load density and interconnection of electrical networks. In addition, OLTCs applied in industrial process transformers as regulating units in the chemical and metallurgical industry is another important field of application [6].

Basic connection diagrams for regulation are typically chosen with regard to system conditions and size/weight limits during transportation.

Linear regulation (Figure 2) is applied for a simple design transformer and regulation unit, with regulating range up to 20 percent of nominal voltage. The taps are added or substracted in the series with the main winding. This kind of regulation ensures smallest ohmic losses.

Figure 2: Basic connection diagram linear regulation

For linear regulation, voltage accros tap winding U_{TV} is equal phase to phase voltage across regulation U_R.

$$U_{TV} = U_R \tag{1}$$

Most common type of regulation is a reversing change-over selector (plus/minus switching), which allows doubled tapping range. Regulated winding is connected with the basic winding in series (Figure 3). The total number of taps decreases or increases, depending on position of tap selector. Boost and buck connection (boost - vectorial addition to main winding and buck -vectorial substraction to main winding) enables to increase the regulating range or to reduce the number of tapped winding.

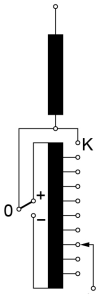


Figure 3: Basic connection diagram with reversing switch

Regulation with reversing change-over selector, for constant induction, depends on regulating voltage and number of step of the tap winding:

$$U_{TV} = \frac{U_R}{2} \cdot \frac{m}{m-1} \quad (2)$$

Coarse change over selector or coarse/fine switching requires more complicated winding layout – it is regulation with two windings, regulated winding and coarse winding insertion (Figure 4). Arrangement of coarse/fine selector has the electrical length of the fine tap winding plus one step. This kind of regulation offers the lower copper losses in the tap position with the mínimum number of effective turns. Coarse/fine regulation offers for some industrial applications possibility of large number of operations using special design with up to 5 coarse taps (107 operations i.e. operating positions) [7].

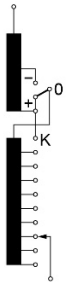


Figure 4: Basic connection diagram with coarse tap selector

Regulation with coarse fine change over selector across tap winding:

$$U_{TV} = \frac{U_R}{2} \quad (3)$$

Voltage across coarse tap winding U_{CV}=U_{TV}

where:

U_{TV} is voltage across tap winding in kV;

U_R is regulating range phase to phase in kV;

U_{CV} is voltage across coarse winding in kV;

m is number of steps of tap winding

3. Manual Selection and Calculations

The selection of appropriate OLTC should be made very carefully, because the OLTC is an important factor with respect to the transformer's reliability and cost. Each type of regulation unit is available in a number of variants, with different values of maximum rated through-current, number of phases, highest insulation level, tap selector size and basic connection diagrams. Therefore, the type designation of a regulation unit depends on these features. A guideline for the OLTC selection is given in IEC 60542, but some special applications as HVDC transformers or phase shifters are not described in detail.

When selecting the appropriate type of OLTC or DETC manually, a number of parameters and technical limits have to be considered.

First of all, the rated through current of the regulation unit must not be less than the transformer highest current value of the regulating winding. For regulation at the neutral end with constant induction (constant flux voltage variation, CFVV), the rated through current in case of a three-phase full transformer is derived using the rated power as the product of current and voltage (one or three-phase).

For a three-phase autotransformer with regulation in neutral end, variable induction (variable flux voltage variation, VFVV),. Sn=const., the rated through current depends on both regulated and non-regulated voltage levels:

$$I_{max} = \frac{S_n \times 1000}{\sqrt{3}} \times \left(\frac{1}{U_{2min}} - \frac{1}{U_1}\right) \quad (4)$$

where:

U_1 is the rated line voltage of the non-regulated side in kV

U_{2min} is the rated line voltage in the minimum tap position in kV

In case of **one-phase transformer**, the rated through current is multiplied by a factor of $\sqrt{3}$ since the rated power is per phase and the rated voltages are phase voltages.

Maximum rated through current is the current which the regulation unit is transferring from one tap to the other at relevant step voltage.

Step voltage U_{step} [V] is the phase voltage between the taps which depends on the required regulated voltage range and the required number of positions. This value is constant in case of CFVV (constant flux voltage variation), unlike in the case of VFVV (variable flux voltage variation) where different regulating positions have different volt/turn ratio and therefore different step voltage values:

$$U_{step} = \begin{cases} \frac{U_{1max} - U_{1min}}{\sqrt{3}(n-1)} \times 1000 & ; \text{CFVV regulation} \\ \max(U_k - U_{k-1}) & ; \text{VFVV regulation} \end{cases} \quad (5)$$

where:

U_{1max} and U_{1min} are the required maximum and minimum regulated line voltage in [kV] (equivalent three-phase bank line voltage in case of a single-phase transformer);

n is the required number of regulating positions (including the rated position);

U_k-U_{k-1} is the (phase) voltage difference in [V] in the regulating position *k* and *k-1* (for positions $k = 2$ to *n*).

The required regulation unit external insulation level depends on the insulation level to ground of the side where the regulation unit is connected. The withstand voltages of the external insulation are standardized by international standards and correspond to the highest voltage for equipment [8]. In case of regulation in neutral end (common for full transformer or VFVV autotransformer), the insulation level is usually the same as the insulation level of the transformer neutral (unless otherwise specified). In case of regulation in line (common for CFVV autotransformers), the insulation level is the same as the lower voltage side to which the regulation unit is connected. The insulation level has to be checked against the applicable technical guide.

The internal insulation level affects the regulation unit tap selector size, whose price can vary significantly, meaning that the proper selection is of the essence in choosing the optimal regulation unit. Highest operating voltage across the regulating winding is the value that has to be compared with the highest

permissible phase service voltage across the regulating winding described in the technical guide, which defines several internal insulation distances to be checked, most important of which are distances 'a' and 'b'. For insulating distance 'a' between start and end of the regulating winding and insulating distance 'b' between the fine tap selector contacts of different phases, a statistics-based estimation is used. The logic behind the estimation is assuming linear distribution, and additionally applying a „non-linearity" factor depending on the regulation concept (reg.in main / coarse-fine / reversing / linear) and transformer type (full / auto). The mentioned factor is determined using statistical analysis derived from experimental data. The differences in factors arise from the specifics of the distribution of electrical field in various power transformer types.

With some winding arrangements (neutral-end in autotransformers and line-end regulation in delta connected windings), very high voltages may occur. These voltages are significantly influenced by the choice of the regulation concept (linear, coarse fine or reversing regulation). In such cases, additional tools are used to determine the electrical field distribution in the power transformer (FEM analysis). Since the tap selector size price varies significantly, this can also be subject to optimization.

The switching capacity P_{StN} [kVA] determines the capability of the regulation unit to switch between regulation steps. Since the worst case has to be considered, the required switching capacity in the regulation unit is the product of the relevant step voltage and maximum rated through current (both described above):

$$P_{StN} = \frac{I_{max} \times U_{step}}{1000} \tag{6}$$

The required switching capacity for a specific contact in an OLTC is based on the relevant step voltage and current but is also determined by the design and circuit of the OLTC. The switching capacity itself is primarily a function of the contact design, contact speed and arc-quenching agent [9].

The selected regulation unit must be within the switching capacity diagram (example in Figure 5). Limit values are given in the manufacturer's technical guide and may rectrict certain types of tap changers in some cases.

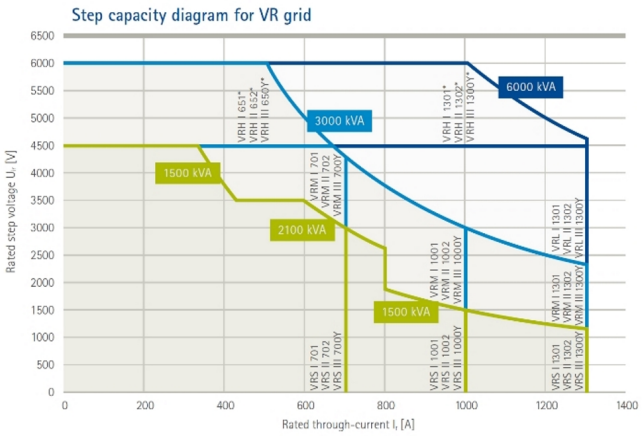


Figure 5: Step capacities diagram

The number of tap positions and the number of contacts have to be considered as well - different types have different maximum regulation range, depending on the regulation concept. A higher number of positions may also increase the voltage stress of the regulating winding, so this has to be taken into consideration (as described in the internal voltage calculations above). For rectifier or furnace transformers, wide tapping range should be considered.

The predicted contact life of fixed and moving contacts is a function of the rated through current and has to be compared with the diagram for a specific regulation unit which can generally be found in the regulation unit manufacturer's technical guide.

The recovery voltage occurs on the open change-over selector contacts during switching sequence. It depends on the geometric winding arrangement and the capacitances C1 and C2. Recovery voltage can cause switching sparks or low-energy arcs, thereby producing unwanted gas which has a negative effect on the characteristics of the insulation media (usually insulating oil). The recovery voltage and breaking current values have to be checked in + and – selector switch positions:

$$|\overrightarrow{U_{W+}}| = \sqrt{\left(\frac{U_1}{2} + \frac{U_f}{2}\right)^2 + \left(\frac{U_1}{2\sqrt{3}} \times \frac{C_2}{C_1 + C_2}\right)^2} \tag{7}$$

$$|\overrightarrow{U_{W-}}| = \sqrt{\left(\frac{U_1}{2} - \frac{U_f}{2}\right)^2 + \left(\frac{U_1}{2\sqrt{3}} \times \frac{C_2}{C_1 + C_2}\right)^2} \tag{8}$$

The breaking current I_{S+} and I_{S-}:

$$\overrightarrow{I_{S+}} = \frac{\overrightarrow{U_1}}{2\sqrt{3}} \times \omega C_2 + j\frac{\overrightarrow{U_1} + \overrightarrow{U_f}}{2} \times \omega(C_1 + C_2) \tag{9}$$

$$\overrightarrow{I_{S-}} = \frac{\overrightarrow{U_1}}{2\sqrt{3}} \times \omega C_2 + j\frac{\overrightarrow{U_1} - \overrightarrow{U_f}}{2} \times \omega(C_1 + C_2) \tag{10}$$

U_{W+} and U_{W-} are the recovery voltages in [kV];

I_{S+} and I_{S-} are breaking currents in [mA];

C_1 and C_2 are the capacitance values in [pF] of the regulating winding (as shown in Figure 6) which can be approximated assuming cylindrical winding system [10].

U_1 is the rated regulated voltage in [kV];

U_f is the preselector voltage in [kV];

$\omega = 2\pi f$, where f is the power frequency in [Hz].

Example for recovery voltage calculation for single phase autotransformer with regulation of common winding (Figure 7) in wye connection (fine regulation), input data:

- Rated power: 200MVA
- Rated voltages and relevant tapping range: 400 / 230 +10 x 1,5%, -10 x 1,5% / 24 kV
- Step voltage: 1992 V
- Max. tapping current: 1772 A

Test Voltages:

- a...340 kV (1,2/50μs); 65 kV (50 Hz)
- U_m... 950 kV (1,2/50μs); 395 kV (50 Hz)

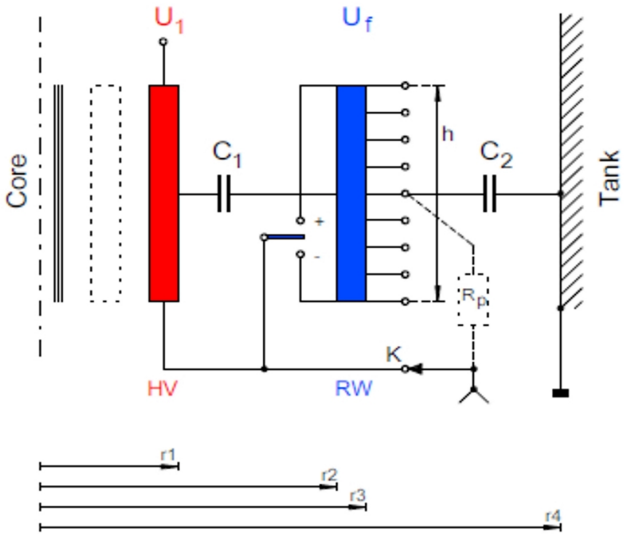


Figure 6: Recovery voltage – capacitances per phase

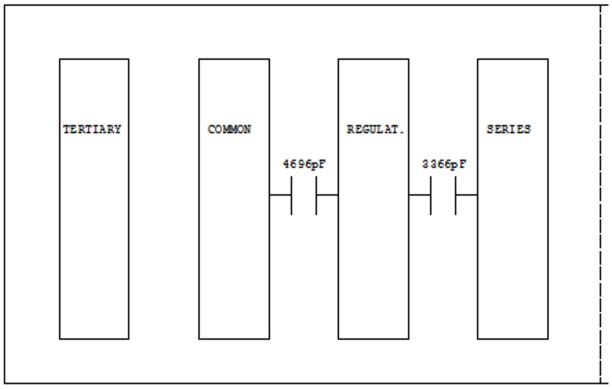


Figure 7: Autotransformer winding arrangement wye connection

According to input data (capacitances C_1=4696pF and C_2=3366pF), following recovery voltages and breaking current are calculated:

U_{W+}=8,3kV; U_{W-}=28,2kV

I_{S+}=21mA; I_{S-}=71,3mA

According to Figure 8., the potential measure is not needed.

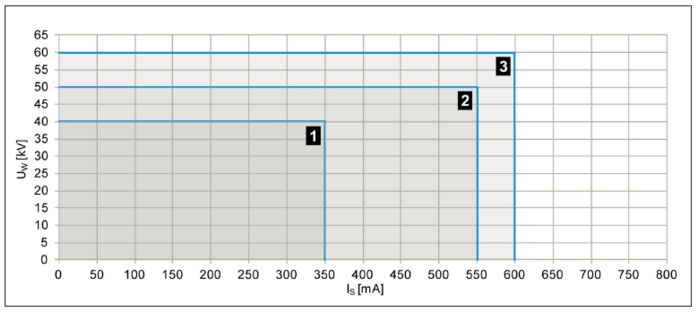


Figure 8: Recovery voltage and breaking current diagram

The recovery voltage permisible value is the function of the calculated breaking current value. The mentioned $U_W(I_S)$ function for each tap selector type and size can be found in the technical data of the regulation unit. If the calculated values of recovery voltages are exceeded, it is necessary to install tie-in resistors or select a different tap selector. The installation of the tie-in resistors affects the regulation unit price as well as overall transformer dimensions, so a solution with bigger tap selector size and therefore higher permisible values of recovery voltage should be considered in such cases. Furthermore, the tie-in resistor dimensions are defined by the regulation unit manufacturer specifically for each design, which can cause an additional unwanted time delay in the transformer final design process. For phase shifting transformers, using regulation unit with two-way change-over selector, one can overcome problems during the change-over selector operation and change of winding connections between phases of the system voltage [11].

4. Tap Changer Selection Application and Database

Due to the complexity of the regulation unit technical limits and cross-dependencies between different types, a relational database is developed in such an architecture that an algorithm can quickly go though all of the possible permutations and compare the actual calculated values with every combination of regulation unit's limits. Also, the database management user interface (UI) is developed in MS Access environment. Within the database, SQL is used for querying data and sending it to the application for further analysis in the TCS algorithm (Tap Changer Selection). The algorithm then compares the queried data with the calculation results which are calculated in the background as thoroughly described in chapter III. The algorithm also takes into account different possibilities of transformer regulation: for example, a three phase transformer can either be regulated with a single three-phase regulation unit or three single-phase regulation units. Furthermore, autotransformer can be designed with regulation in main winding, meaning that the regulation unit does not have a common neutral end. Therefore, autotransformers with such regulation can only be regulated with three single-phase regulation units (three-phase regulation unit requires a neutral end common point). Also, different regulation concepts have different possibilities for number of positions and number of contacts.

All of the mentioned above is automatically checked for every single regulation unit combination queried from the database. This is achieved using a foreach loop which (in each step) calls for different modules used for this purpose. All of the modules and functions are developed in C#, following object-oriented software development principles (Model-View-Controller, MVC pattern). The input for the algorithm that is expected from the application user is the same data that the electrical designer uses in power transformer design (MVA rating, rated voltages and required regulation range, regulation concept, connection of regulating winding, LI and AC voltage levels i.e. external insulation levels to ground and number of tap positions), making the application very user-friendly for any power transformer designer. This data model also makes automatic data import from transformer design software very simple. The algorithm output is the list of all the possible solutions for regulation units (tap changers) that satisfy the given input.

The results are finally sorted according to the expected total cost, as shown in Figure 9. Therefore, the application user can simply choose the first regulation unit from the generated list knowing that this is the optimal solution. Other regulation units from the list are not optimal but are also technically acceptable. These can be selected in case of a specific request (e.g. higher

external insulation level requested by the customer). Finally, a report containing all the calculations together with the technical limits can be generated automatically.

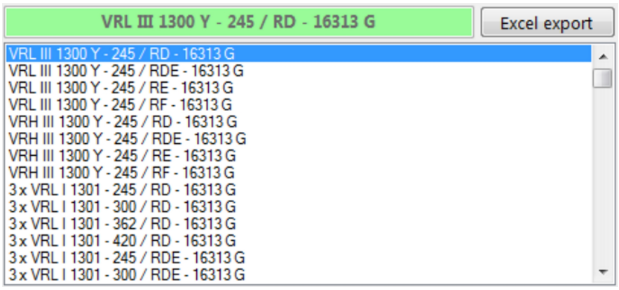


Figure 9: TCS results

4.1. Web application

Together with the Tap Changer Selection application, an additional tool (Figure 10) is developed for accessing the final order data through a Web interface.

The *RSB Base* application allows reading all the necessary technical data from the ordering data sheets, saves this data to the database and gives the possibility to search, analyze and report through user-friendly interface.

Features and functions of the MVC WEB application and the database:

- Tables
  OLTC data loaded from ordering documenation, gets entered into the database tables. An intuitive UI provides an organized way to represent the data. Tables contain all necessary fields from ordering documentation as well as coded tables for data analyzing and reporting.
- Forms
  Application provide functionality with controls that allow inserting new ordering data, editing, deleting, page through, sort, filter and search tables and queries, modify ordering data, as well as creating reports.
- Reports
  The application allows the user to manipulate stored data in various ways. One can use application's built-in function for creating reports to help in analyzing the OLTC data, e.g., calculation of motor drive equipment and electronics in detail.
- Data functions
  Application software provides features to organize data (CRUD - Create Read Update and Delete) and functions for sorting, filtering, reporting etc.

5. Results and Future Work

Possible future work includes improvements by developing a faster OLTC search algorithm. The database contains all the necessary technical information about the offered and executed projects. The current available size is approximately 400 items with more than 90 properties (with a growth tendency). By increasing database and displaying equipment in the motor drive unit in detail, the possibility to make a mistake in the transformer design offer stage is already significantly reduced. In TCS application, different simplifying estimations were made (depending on the regulation concept, linear distribution and nonlinear factor were applied) which are still acurate enough to select an OLTC properly. There is room for expansion in the application in order to better cover some special transformer's designs, such as HVDC, limit cases or arrangement with enforced current splitting. However, the results are already visible in the offering phase, by reducing the time of OLTC selection up to 90% compared to the manual selection. This means a lot of working hours in the design office have been saved. Further improvements are reflected in the pooling data from application database itself and from the web application that is related to the OLTC and motor drive unit equipment ordering data in one general database.

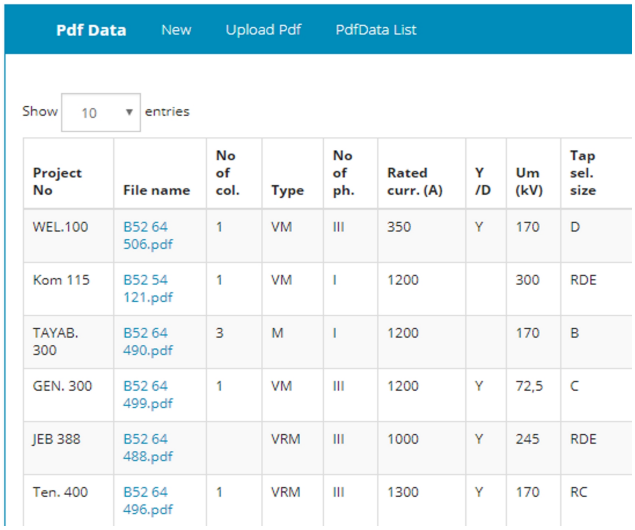


Figure 10: Web interface – homescreen

6. Conclusion

The importance of regulation in power transformers shows that optimizing and choosing the appropriate regulation unit may have a significant impact on the transformer overall cost, making it one of the essential parts of power transformer design. Management of such a complex system by implementing a software solution is presented in this paper. Comparison with Compas tool is not entirely possible, because TCS application works with types of OLTCs and DETCs other than ABB. It could be said that TCS and Compas largely cover all the needs of the transformer manufacturer for the appropriate regulation unit. In TCS application, dozens of tables and diagrams, as well as hundreds of pages of technical data from different manufacturers are incorporated in the database. Nevertheless, the developed algorithm performs all the calculations and searches through all of the possibilities in a matter of seconds, regardless of the regulation unit manufacturer. This represents significant time savings and reduced possibility of errors, providing productivity improvement in power transformer design process. Selection of OLTC is a critical point in the transformer design process due to a large number of correct solutions, only one of them being optimal while also technicaly acceptable and cost-effective. It can be concluded that the TCS application connects the physical properties of the OLTC and the information technology, which has led to material and time savings in transformer design process

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M.Gazdović, T.Šimović, "Digitalization of Regulation Unit Selection Process in Power Transformer Design" Proceedings of the 41st International Convention for Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018).

[2] G.Leci, A.Marušić "Koordinirana regulacija napona energetskih transformatora" 11. savjetovanje HRO CIGRÉ, 2013.

[3] IEC 60214-2 Tap-changers – Part 2: Application guide, 2004

[4] ABB, Tap-changer selection program Compas [Web] http://new.abb.com/products/transformers/interactive-tools/tap-changer-selection-program-compas

[5] R.Levi, "On Load Tap Changer Condition Assessment Using Dynamic Recording and Measurement (DRM)", 2017.

[6] A.Krämer, "On-load tap-changers for power transformers," Maschinenfabrik Reinhausen, 2000.

[7] Maschinenfabrik Reinhausen GmbH "TD - General Section", 2003.

[8] IEC 60076 – 3 Power tansformers – Part 3: "Insulation levels, dielectric tests and external clearances in air", 2013.

[9] D.Dohnal, "On-Load Tap-Changers for power transformers," Maschinenfabrik Reinhausen, 2013.

[10] M.Gazdović, "TCS help", Končar Power Transformers, 2017., unpublished.

[11] T.Šimović, M.Stanić, L.Kirchner, "200 MVA phase shifting autotransformer for HE Senj and the possibility to change-over between active and reactive power flow control in energized state", 4rd International Colloquium "Transformer Research and Asset Management", May 10-12, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 8-16 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Robust synchronization of nonfragile control of complex dynamical network with stochastic coupling and time-varying delays

Fehrs Adu-Gyamfi[*,1,2], Yuhua Cheng[1], Chun Yin[1]

[1]*University of Electronic Science and Technology of China (UESTC), School of Automation,611731, Chengdu - P.R. China*

[2]*Koforidua Technical University (KTU), Mechanical Engineering Department, Faculty of Engineering, KF 981, Koforidua-Ghana*

ARTICLEINFO

Article history:
Received: 04 December, 2018
Accepted: 22 February, 2019
Online: 10 March, 2019

Keywords:
Complex dynamical networks
Extended Jensen's integral inequality
Nonfragile control
Robust
Stochastic coupling
Synchronization

ABSTRACT

This paper explores the problem of robust synchronization of complex dynamical network with stochastic coupling and time-varying delays through the application of nonfragile control. A well defined Lyapunov Krasovskii functional is established and by employing the widely acknowledged extended Jensen's integral inequality and the Bernoulli's distribution sequences, the stochastic nature of network coupling is modeled which entails the occurrence of randomness in the controller gain uncertainties are presented. Sufficient delay dependent conditions are given for the purposes of synchronization. Additionally, a nonfragile controller is designed based on linear matrix inequalities (LMIs). Two numerical examples are finally given to exhibit the effectiveness and usefulness of the proposed theoretical results.

1 Introduction

This article is an extension of a paper previously presented in the international conference of ubiquitous and future networks 2018 (ICUFN 2018) [1]. The great works of Watts and Strogatz [2] which focused on investigation of complex networks have witnessed a tremendous attention from many scientific communities because of the theoretical importance and practical implementation of such outcome in areas such as computer networks, social networks, biological networks, communication networks, electric power grid, food webs and transportation networks [3, 4, 5, 6, 7, 8]. Complex dynamical networks (CDNs) are large number of interconnected nodes with each node having some defined contents. Majority of these networks display some level of complexities in their overall topology as well as dynamical properties [9]. Amongst the important collective behaviors of CDNs is synchronization problem. This behavior has been investigated by many profound researchers [10, 11]. Synchronization of a network is related to subsystems been represented as nodes in coupled systems in which the various nodes with different initial conditions converge to a common behavioral trajectory. Many control methodologies have been proposed to ensure the solution to synchronization problems, amongst such control methods are pinning control [12], sampled data control [13], sliding mode control [14], impulsive control [15] and so on. The tendencies of sudden changes in network coupling which can emanate from internal and external environmental factors such as unexpected change of working environment, random link failures and repairs on network connectivity result in stochastic behavior in the network coupling. Additionally, the practical implementation of control design might not be precise because of limited information speed, round-off error of numerical computations, aging of system components, analogy to digital conversions (ADC) and Digital to analogy conversion (DAC) [13]. Controller fragility refers to the variation effects on the control parameters as cited in [16]. This problem is addressed in the design of a nonfragile control

[*]Corresponding Author: Fehrs Adu-Gyamfi, UESTC and KTU, Email: fehrs.adugyamfi@ktu.edu.gh

https://dx.doi.org/10.25046/aj040202

scheme which is presented to address controller gain fluctuation based on the aforementioned shortfalls as presented in [17, 18].
Time delays which are ubiquitous in nature and present in CDNs, possess the ability to destroy synchronization performance which can lead to oscillation and instability of the network, hence the need to consider it when addressing synchronization issues of CDNs. It is very important to also note that, there exit time delays in the exchange of information resulting from finite transmission speed in the network, memory effect, limited bandwidth and so on [19]. From the aforementioned discussions, we are motivated in this paper to seek solutions to CDNs synchronization with nonfragile control design scheme, taking into consideration the stochastic coupling nature of the network. The important contributions of this paper are summarized as follows:

1. The problem of CDNs robust synchronization with stochastic coupling and time-varying delays is studied permitting some level of control gain uncertainties.
2. Control scheme with non-fragile characteristics is designed and presented which guarantee the system error synchronization.
3. A suitable Lyapunov Krasovskii functional (LKF) is chosen which applied the extended Jensen's integral inequality.
4. Given are two numerical examples which indicate the usefulness of our proposed control approach.

2 Model description and Preliminaries

In this paper, standard notations are used. $\mathbb{R}^n$ shows the Euclidean n- dimensional space, $\mathbb{R}^{m\times n}$ denotes the set of all $m\times n$ real matrices. I and 0 represent identity and zero matrices with appropriate dimensions, respectively. $P>0$ is a real positive symmetric definite matrix. The superscript "T" indicates transposition. Also, an asterisk$(*)$ is used to show the symmetric terms and diag{....} represent a diagonal block matrix. All other matrices without given dimensions are considered to be of compatible dimensions. In this paper, the robust synchronization of CDNs which entail the stochastic coupling of N identical nodes with time varying delays is described as follows:

$$\begin{aligned}\dot{\tilde{r}}_i(t) &= A\tilde{r}_i(t)+Bf(\tilde{r}_i(t))+(1-\bar{\delta}_1(t))\sum_{j=1}^{N}\bar{c}_{ij}\Gamma\tilde{r}_j(t)\\ &+\bar{\delta}_1(t)\sum_{j=1}^{N}\bar{c}_{ij}\Gamma\tilde{r}_j(t-\ddot{\gamma}(t))+\hat{u}_i(t),\\ &i=1,2,\ldots,N,\end{aligned}\tag{1}$$

Where, $\tilde{r}_i(t)=(\tilde{r}_{i1}(t),\tilde{r}_{i2}(t),...,\tilde{r}_{in}(t))^T\in\mathbb{R}^n$ denotes the state vector of the i^{th} node, $\hat{u}_i(t)\in\mathbb{R}^n$. A,B are known real constant matrices and they are assumed to be stabilizable, $f:\mathbb{R}^n\to\mathbb{R}^n$ indicate a smooth nonlinear function, $\Gamma\in\mathbb{R}^{n\times n}$ is the inner coupling matrix of the nodes and $\bar{C}=(\bar{c}_{ij})_{N\times N}$ represent the outer coupling configuration of the network. If there is a connection from node i to node j $(i\neq j)$, then the coupling matrix $\bar{c}_{ij}\neq 0$; otherwise $\bar{c}_{ij}=0$. Furthermore, the diagonal elements are defined as $\bar{c}_{ii}=-\sum_{j=1,j\neq i}^{N}\bar{c}_{ij}$. $\ddot{\gamma}(t)$ represent time-varying delay which is considered to be a differentiable function that satisfies the following conditions:

$$0\leq\ddot{\gamma}_1\leq\ddot{\gamma}(t)\leq\ddot{\gamma}_2,\qquad\dot{\ddot{\gamma}}(t)\leq\tilde{\mu}.\tag{2}$$

$\bar{\delta}_1(t)\in\mathbb{R}$ denotes a stochastic variable, which is in the form of a Bernoulli distribution sequence defined by

$$\bar{\delta}_1(t)=\begin{cases}1 & \text{presence of delay in information exchange,}\\ 0 & \text{no delay in information exchanges}\end{cases}$$

Below indicates the stochastic probability variable $\bar{\delta}_1(t)$:

$$\begin{aligned}Pr\{\bar{\delta}_1(t)=1\}&=\bar{\delta}_1,\\ Pr\{\bar{\delta}_1(t)=0\}&=1-\bar{\delta}_1,\end{aligned}$$

where $\bar{\delta}_1\in[0,1]$ is a known constant.
Then, the supposed initial condition for (1) is given by $\tilde{r}_i(t)=\bar{\psi}_i(t)$, $t\in[-\ddot{\gamma}_2,0]$ and $i=1,\ldots,N$.

Assumption 2.1 *[20] The continuous vector valued function $f(\bullet):\mathbb{R}^n\to\mathbb{R}^n$ is considered which satisfies $f(0)=0$, hence this sector-bounded condition stands:*

$$[f(x)-f(y)-Z_1(x-y)]^T[f(x)-f(y)-Z_2(x-y)]\leq 0,\tag{3}$$

given that Z_1 and Z_2 are constant matrices of appropriate size. From (1) and using kronecker properties, we have

$$\begin{aligned}\dot{\tilde{r}}(t)&=(I_N\otimes A)\tilde{r}(t)+(I_N\otimes B)g(\tilde{r}(t))+(1-\bar{\delta}_1(t))(\bar{C}\otimes\Gamma)\\ &\times\tilde{r}(t)+\bar{\delta}_1(t)(\bar{C}\otimes\Gamma)\tilde{r}(t-\ddot{\gamma}(t))+\hat{u}(t)\end{aligned}$$

Where,

$$\begin{aligned}\tilde{r}(t)&=[\tilde{r}_1^T(t),\tilde{r}_2^T(t),\ldots,\tilde{r}_N^T(t)]^T,\\ g(\tilde{r}(t))&=[f^T(\tilde{r}_1(t)),f^T(\tilde{r}_2(t)),\ldots,f^T(\tilde{r}_N(t))]^T,\\ \hat{u}(t)&=[\hat{u}_1^T(t),\hat{u}_2^T(t),\ldots,\hat{u}_N^T(t)]^T.\end{aligned}$$

Lemma 2.1 *([21] Jensen inequality) Given a matrix $H=H^T>0$, of an appropriate dimension and a vector function $\bar{\alpha}(\cdot):[0\quad\bar{\gamma}]\to\mathbb{R}^n$ for a scalar $\bar{\gamma}>0$, the integration is defined as follows:*

$$\bar{\gamma}\int_0^{\bar{\gamma}}\bar{\alpha}^T(s)H\bar{\alpha}(s)ds\geq\left[\int_0^{\bar{\gamma}}\bar{\alpha}(s)ds\right]^T H\left[\int_0^{\bar{\gamma}}\bar{\alpha}(s)ds\right]\tag{4}$$

Lemma 2.2 *[22] For constant matrices* $L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}$, $L \in \mathbb{R}^{2n\times 2n}$, $\tilde{U} = \tilde{U}^T > 0$, $\tilde{U} \in \mathbb{R}^{n\times n}$, *a continuous function satisfying* $d_1 \le d(t) \le d_2$, *and a continuously differentiable function* $x : [-d_2, 0] \to \mathbb{R}^n$ *such that the integration is properly defined, hence the following inequality holds:* $\int_{t-d_2}^{t-d_1} \dot{x}^T(s)\tilde{U}\dot{x}(s)ds \ge \frac{1}{d_{12}} v^T(t)\Psi v(t)$

where

$$
\begin{aligned}
v(t) &= [\quad v_1^T(t) \quad v_2^T(t) \quad v_3^T(t) \quad v_4^T(t) \quad]^T \\
d_{12} &= d_2 - d_1 \\
v_1(t) &= x(t-d_1) - x(t-d(t)) \\
v_2(t) &= x(t-d_1) + x(t-d(t)) - \frac{2}{d(t)-d_1}\int_{t-d(t)}^{t-d_1} x(s)ds \\
v_3(t) &= x(t-d(t)) - x(t-d_2) \\
v_4(t) &= x(t-d(t)) + x((t-d_2) - \frac{2}{d_2-d(t)}\int_{t-d_2}^{t-d(t)} x(s)ds \\
\Psi &= \begin{bmatrix} \tilde{U} & 0 & L_{11} & L_{12} \\ * & 3\tilde{U} & L_{21} & L_{22} \\ * & * & \tilde{U} & 0 \\ * & * & * & 3\tilde{U} \end{bmatrix} \ge 0.
\end{aligned}
\tag{5}
$$

Remark 1. The Lemma (2.2) is derived from the reciprocal convex combination techniques with Jensen inequality resulting in the less conservativeness of our results. The detailed proof is omitted but can be referred from [22].

Lemma 2.3 *(Schur Complement [23]): For a given matix*

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ * & \Lambda_{22} \end{pmatrix} < 0,$$

any of the inequalities below is equivalent to Λ *:*

1. $\Lambda_{11} < 0, \quad \Lambda_{22} - \Lambda_{12}^T\Lambda_{11}^{-1}\Lambda_{12} < 0$
2. $\Lambda_{22} < 0, \quad \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{12}^T < 0$

Lemma 2.4 *[6], Suppose* $Q = Q^T$, R *and* T *been real appropriate dimension matrices and the function* $F(t)$ *which satisfies the condition* $F^T(t)F(t) < I$. *Accordingly,* $Q + RF(t)T + T^TF^T(t)R^T < 0$ *when a given scalar* $\epsilon > 0$ *exist. Then,*

$$\begin{bmatrix} Q & R & \epsilon T^T \\ * & -\epsilon I & 0 \\ * & * & -\epsilon I \end{bmatrix} < 0.$$

Definition 1 *The CDNs* (1) *is synchronized when the condition below holds:*

$$\lim_{t\to\infty}[\tilde{r}_i(t) - s(t)] = 0.$$

Consider the dynamics of an isolated unforced node $s(t)$ to be $\dot{s}(t) = As(t) + Bf(s((t)))$. $s(t)$ can be taken as an equilibrium point, periodic orbit, or even a chaotic attractor.

Let the i^{th} node error system be given as $e_i(t) = \tilde{r}_i(t) - s(t)$. Consequently, error dynamics of CDNs (1) is given by:

$$
\begin{aligned}
\dot{e}_i(t) = Ae_i(t) + Bg(e_i(t)) + (1-\bar{\delta}_1(t))\sum_{j=1}^{N}\bar{c}_{ij}\Gamma e_j(t) \\
+ \bar{\delta}_1(t)\sum_{j=1}^{N}\bar{c}_{ij}\Gamma e_j(t-\ddot{\gamma}(t)) + \hat{u}_i(t)
\end{aligned}
\tag{6}
$$

Note: $g(e_i(t)) \equiv f(\tilde{r}_i(t)) - f(s(t))$.

The designed control scheme to guarantee synchronization of the CDNs is:

$$\hat{u}_i(t) = (k_i + \sigma(t)\triangle k_i(t))e(t) + k_\tau e(t-\ddot{\gamma}(t)), \quad i = 1, 2, \ldots, N. \tag{7}$$

Consider $k_i, k_\tau \in \mathbb{R}^{n\times n}$ as the feedback controller gain matrices which are yet to be estimated. Also, $\triangle k_i$ represent the gain fluctuation, where $\triangle k_i(t)$ is known as follows:

$$\triangle k_i(t) \equiv H_i\Upsilon_i(t)W_i$$

where $\Upsilon_i(t) \in \mathbb{R}^{k\times l}$, is an unknown time-varying matrix which satisfies the condition: $\Upsilon_i(t)^T\Upsilon_i(t) \le I$, H_i and W_i are matrices of known parameters. The probability variable $\sigma(t)$ shows controller gains fluctuations. The random occurring fluctuations of the gain obeys the Bernoulli distribution with the following definition

$$\sigma(t) = \begin{cases} 1 & \text{Gain fluctuation occur,} \\ 0 & \text{Gain fluctuation does not occur} \end{cases}$$

and the probability of stochastic parameter $\sigma(t)$ is given as:

$$
\begin{aligned}
Pr\{\sigma(t) = 1\} &= \sigma \\
Pr\{\sigma(t) = 0\} &= 1-\sigma; \quad \sigma \in [0, 1]
\end{aligned}
$$

Closed loop of the error dynamics (6) of the CDNs yields:

$$
\begin{aligned}
\dot{e}_i(t) &= Ae_i(t) + Bg(e_i(t)) + (1-\bar{\delta}_1(t))\sum_{j=1}^{N}\bar{c}_{ij}\Gamma e_j(t) \\
&\quad + \bar{\delta}_1(t)\sum_{j=1}^{N}\bar{c}_{ij}\Gamma e_j(t-\ddot{\gamma}(t)) + (k_i + \sigma(t)\triangle k_i(t))e_i(t) \\
&\quad + k_\tau e_i(t-\gamma(t)) \\
&= (A + k_i + \sigma(t)\triangle k_i(t))e_i(t) + Bg(e_i(t)) \\
&\quad + (1-\bar{\delta}_1(t))\sum_{j=1}^{N}\bar{c}_{ij}\Gamma e_j(t) + \bar{\delta}_1(t)\sum_{j=1}^{N}\bar{c}_{ij}\Gamma \\
&\quad \times e_j(t-\ddot{\gamma}(t)) + k_\tau e_i(t-\ddot{\gamma}(t))
\end{aligned}
$$

$$
\begin{aligned}
&= (A + k + \sigma H_i \Upsilon_i W_i) e_i(t) + ((\sigma(t) - \sigma) H_i \Upsilon_i W_i) e_i(t) \\
&\quad + Bg(e_i(t)) + (1 - \bar{\delta}_1) \sum_{j=1}^{N} \bar{c}_{ij} \Gamma e_j(t) + (\bar{\delta}_1 - \bar{\delta}_1(t)) \sum_{j=1}^{N} \\
&\quad \times \bar{c}_{ij} \Gamma e_j(t) + \bar{\delta}_1 \sum_{j=1}^{N} \bar{c}_{ij} \Gamma e_j(t - \ddot{\gamma}(t)) + (\bar{\delta}_1(t) - \bar{\delta}_1) \\
&\quad \times \sum_{j=1}^{N} \bar{c}_{ij} \Gamma e_j(t - \ddot{\gamma}(t)) \\
&\quad + k_\tau e_i(t - \ddot{\gamma}(t)); \quad (i = 1, 2, \ldots, N)
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
\dot{e}(t) &= (\bar{A} + \bar{k} + \sigma \bar{H} \bar{\Upsilon} \bar{W}) e(t) + (\sigma(t) - \sigma) \bar{H} \bar{\Upsilon} \bar{W} e(t) \\
&\quad + \bar{B} G(e(t)) + (1 - \bar{\delta}_1)(\bar{C} \otimes \Gamma) e(t) + (\bar{\delta}_1 - \bar{\delta}_1(t)) \\
&\quad \times (\bar{C} \otimes \Gamma) e(t) + \bar{\delta}_1 (\bar{C} \otimes \Gamma)(e(t) - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s) \\
&\quad + (\bar{\delta}_1(t) - \bar{\delta}_1)(\bar{C} \otimes \Gamma)(e(t) - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s) \\
&\quad + \bar{k}_\tau (e(t) - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s)
\end{aligned} \tag{9}
$$

where,
$\bar{A} \equiv I_N \otimes A$
$\bar{k} \equiv diag(k_1, k_2, \ldots, k_N)$
$\bar{H} \equiv diag(H_1, H_2, \ldots, H_N)$
$\bar{\Upsilon} \equiv diag(\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_N)$
$\bar{W} \equiv diag(W_1, W_2, \ldots, W_N)$
$e(t) \equiv [e_1^T(t), e_2^T(t), \ldots, e_N^T(t)]^T$
$G(e(t)) \equiv [g^T(e_1(t)), g^T(e_2(t)), \ldots, g^T(e_N(t))]^T$
$\bar{k}_\tau \equiv I_N \otimes k_\tau$
$\bar{B} \equiv I_N \otimes B$

$$
\begin{aligned}
\dot{e}(t) &= (\bar{A} + \bar{k} + \sigma \bar{H} \bar{\Upsilon} \bar{W} + (\bar{C} \otimes \Gamma)) e(t) + ((\sigma(t) - \sigma) \\
&\quad \times \bar{H} \bar{\Upsilon} \bar{W}) e(t) + \bar{B} G(e(t)) - \bar{\delta}_1 (\bar{C} \otimes \Gamma) e(t) + \bar{\delta}_1 (\bar{C} \otimes \Gamma) \\
&\quad \times (e(t) - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s + (\bar{\delta}_1 - \bar{\delta}_1(t))(\bar{C} \otimes \Gamma) e(t) \\
&\quad + (\delta_1(t) - \bar{\delta}_1)(\bar{C} \otimes \Gamma)(e(t) - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s + \bar{k}_\tau (e(t) \\
&\quad - \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s \\
&= (\bar{A} + \bar{k} + \sigma \bar{H} \bar{\Upsilon} \bar{W} + (\bar{C} \otimes \Gamma) + \bar{k}_\tau) e(t) + (\sigma(t) - \sigma) \\
&\quad \times \bar{H} \bar{\Upsilon} \bar{W} e(t) + \bar{B} G(e(t)) - (\bar{k}_\tau + \bar{\delta}_1 (\bar{C} \otimes \Gamma)) \\
&\quad \times \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s - (\bar{\delta}_1(t) - \bar{\delta}_1)(\bar{C} \otimes \Gamma) \int_{t - \ddot{\gamma}(t)}^{t} \dot{e}(s)\, \mathrm{d}s
\end{aligned} \tag{10}
$$

3 Main results

This section establishes sufficient condition for the synchronization purposes of the CDNs. Additionally, the method for designing the synchronization controllers are presented in terms of LMIs.

Theorem 1 *Suppose Assumption 1 holds. For given scalars* $\sigma \in [0,1]$, $\bar{\delta}_1, \ddot{\gamma}_1, \ddot{\gamma}_2$, *and* $\tilde{\mu} < 1$, *the closed-loop error system* (10) *is synchronized with controller gains* $k_i, k_{\tau i} (i = 1, 2, \ldots, N)$. *If some positive definite matrices* $P, \hat{S}, W_1, W_2, Z \in \mathbb{R}^{nN \times nN}$ *exist and any given matrices* $\begin{bmatrix} L_{11} & L_{12} \\ * & L_{22} \end{bmatrix}$ *such that*

$$
\Psi = \begin{bmatrix} \hat{S} & 0 & L_{11} & L_{12} \\ * & 3\hat{S} & L_{21} & L_{22} \\ * & * & \hat{S} & 0 \\ * & * & * & 3\hat{S} \end{bmatrix} \geq 0,
$$

Then

$$
\begin{bmatrix} \Phi_1 & N & \varepsilon W_*^T \\ * & -\varepsilon I & 0 \\ * & * & -\varepsilon I \end{bmatrix} < 0, \tag{11}
$$

holds, as

$$
\Phi_1 = \begin{bmatrix} \Omega_1 & M_{1*}^T P & \sqrt{\sigma(1-\sigma)} M_2^T P & 0 \\ * & \Theta - 2P & 0 & 0 \\ * & * & \Theta - 2P & 0 \\ * & * & * & \Theta - 2P \end{bmatrix}
$$

$$
N^T = \begin{bmatrix} \sigma(P\bar{H})^T & \underbrace{0 \cdots 0}_{8\,elements} & \sigma(P\bar{H})^T & 0 & \hat{\omega} \end{bmatrix}
$$

$$
W_* = \begin{bmatrix} W & \underbrace{0 \cdots 0}_{11\,elements} \end{bmatrix}
$$

$$
M_1 = \begin{bmatrix} \bar{Z} & \bar{B} & -(k_\tau + \bar{\delta}_1(\bar{C} \otimes \Gamma)) & \underbrace{0 \cdots 0}_{6\,elements} \end{bmatrix}
$$

$$
\bar{Z} = \bar{A} + \bar{k} + \sigma \bar{H} \bar{\Upsilon} \bar{W} + \bar{C} \otimes \Gamma + \bar{k}_\tau
$$

$$
M_{1*} = M_1 - \begin{bmatrix} \sigma \bar{H} \bar{\Upsilon} \bar{W} & \underbrace{0 \cdots 0}_{8\,elements} \end{bmatrix},
$$

$$
M_2 = \begin{bmatrix} 0 & 0 & \bar{C} \otimes \Gamma & \underbrace{0 \cdots 0}_{6\,elements} \end{bmatrix}, \Omega_1 \equiv \begin{bmatrix} \Omega_1^1 & \Omega_1^2 \\ * & \Omega_1^3 \end{bmatrix}
$$

$$
-\tfrac{1}{\ddot{\gamma}_{12}} \Pi^T \Psi \Pi, \quad \Omega_1^1 \equiv \begin{bmatrix} ß_1 + ß_2 & P\bar{B} + F_2 \\ * & -I \end{bmatrix},
$$

$$
ß_1 = P\bar{A} + \bar{A}^T P + P\bar{k} + \bar{k}^T P + P(\bar{C} \otimes \Gamma) + (\bar{C} \otimes \Gamma)^T P
$$

$$
ß_2 = P\bar{k}_\tau + \bar{k}_\tau^T P + W_1 + \ddot{\gamma}_1 W_2 - F_1
$$

$$
\Omega_1^2 \equiv \begin{bmatrix} -\bar{\delta}_1 P(\bar{C} \otimes \Gamma) - P\bar{k}_\tau & \underbrace{0 \cdots 0}_{6\,elements} \\ 0 & \underbrace{0 \cdots 0}_{6\,elements} \end{bmatrix},
$$

$$
\Omega_1^3 \equiv \begin{bmatrix} -\frac{1-\tilde{\mu}}{\ddot{\gamma}_2} Z & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \tilde{\Delta}_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{\ddot{\gamma}_1} W_2 \end{bmatrix}
$$

Where:
$\Theta = \ddot{\gamma}_{12} \hat{S} + \ddot{\gamma}_2 Z$, $\tilde{\Delta}_1 = -(1 - \tilde{\mu}) W_1 - F_1$,
$\hat{\omega} = \sqrt{\sigma(1-\sigma)} (P\bar{H})^T$, $F_1 = I_N \otimes \frac{U_1^T U_2 + U_2^T U_1}{2}$,
$F_2 = I_N \otimes \frac{U_1^T + U_2^T}{2}$

Proof. The following candidate of Lyapunov-Krasovskii functional is considered

$$\bar{V}(t) = \bar{V}_1(t) + \bar{V}_2(t) + \bar{V}_3(t) \tag{12}$$

Where

$$\bar{V}_1(t) = e^T(t)Pe(t)$$

$$\bar{V}_2(t) = \int_{t-\ddot{\gamma}(t)}^{t} e^T(s)W_1e(s)\mathrm{d}s + \int_{-\ddot{\gamma}_1}^{0}\int_{t+\phi}^{t} e^T(s)W_2e(s) \times \mathrm{d}s\mathrm{d}\phi$$

$$\bar{V}_3(t) = \int_{-\ddot{\gamma}_2}^{-\ddot{\gamma}_1}\int_{t+\delta}^{t} \dot{e}^T(s)\hat{S}\dot{e}(s)\mathrm{d}s\mathrm{d}\delta \quad + \int_{-\ddot{\gamma}(t)}^{0}\int_{t+\delta}^{t} \dot{e}^T(s) \times Ze(s)\mathrm{d}s\mathrm{d}\delta$$

Finding infinitesimal operator Ł on $\bar{V}(t)$ results in the following:
$Ł\bar{V}(t) = \lim_{\triangle\to 0^+} \frac{1}{\triangle}\{\mathbf{E}\{\bar{V}(t+\triangle)\} - \bar{V}(t)\}$
One should take note of the following expectation:
$\mathbf{E}\{\bar{\delta}_1(t) - \bar{\delta}_1\} = 0, \quad \mathbf{E}\{(\bar{\delta}_1(t) - \bar{\delta}_1)^2\} = \bar{\delta}_1(1-\bar{\delta}_1)$,
$\mathbf{E}\{\sigma(t) - \sigma\} = 0, \quad \mathbf{E}\{(\sigma(t) - \sigma)^2\} = \sigma(1-\sigma)$.
$\dot{\bar{V}}(t)$ is calculated based on the trajectory of error system (10)

$$\begin{aligned}\mathbf{E}\{L\bar{V}_1(t)\} = \mathbf{E}\{2e^T(t)P[(\bar{A} + \bar{k} + \sigma\bar{H}\bar{\Upsilon}\bar{W} + (\bar{C}\otimes\Gamma) \\ + \bar{k}_\tau)e(t) + \bar{B}G(e(t)) - \bar{\delta}_1(\bar{C}\otimes\Gamma)\int_{t-\ddot{\gamma}(t)}^{t}\dot{e}(s)\,\mathrm{d}s \\ - \bar{k}_\tau\int_{t-\ddot{\gamma}(t)}^{t}\dot{e}(s)\,\mathrm{d}s]\}\end{aligned} \tag{13}$$

$$\begin{aligned}\mathbf{E}\{L\bar{V}_2(t)\} \le \mathbf{E}\{e^T(t)(W_1 + \ddot{\gamma}_1W_2)e(t) - (1-\tilde{\mu}) \\ \times e^T(t-\ddot{\gamma}(t))W_1e(t-\ddot{\gamma}(t)) - \frac{1}{\ddot{\gamma}_1} \\ \times\left(\int_{t-\ddot{\gamma}_1}^{t} e(\theta)\,\mathrm{d}\theta\right)^T W_2\left(\int_{t-\ddot{\gamma}_1}^{t} e(\theta)\,\mathrm{d}\theta\right)\}\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{E}\{L\bar{V}_3(t)\} \le \mathbf{E}\{(\ddot{\gamma}_2 - \ddot{\gamma}_1)\dot{e}^T(t)\hat{S}\dot{e}(t) + \ddot{\gamma}_2\dot{e}^T(t)Z\dot{e}(t) \\ - \int_{t-\ddot{\gamma}_2}^{t-\ddot{\gamma}_1}\dot{e}^T(\delta)\hat{S}\dot{e}(\delta)\,\mathrm{d}\delta - (1-\tilde{\mu})\int_{t-\ddot{\gamma}(t)}^{t} \\ \times\dot{e}^T(s)Z\dot{e}(s)\,\mathrm{d}s\}\end{aligned} \tag{15}$$

Based on lemma (2.2) and from (15), the integral component satisfies the following inequality:

$$-\int_{t-\ddot{\gamma}_2}^{t-\ddot{\gamma}_1}\dot{e}^T(\delta)\hat{S}\dot{e}(\delta)\,\mathrm{d}\delta \le -\zeta^T(t)\frac{1}{\ddot{\gamma}_{12}}\Pi^T\Psi\Pi\zeta(t) \tag{16}$$

Where,

$$\ddot{\gamma}_{12} = \ddot{\gamma}_2 - \ddot{\gamma}_1$$

$$\begin{aligned}v(t) &= \begin{pmatrix} \underbrace{0\cdots0}_{3\,elements} & I & -I & 0 & 0 & 0 & 0 \\ \underbrace{0\cdots0}_{3\,elements} & I & I & 0 & -2I & 0 & 0 \\ \underbrace{0\cdots0}_{3\,elements} & 0 & I & -I & 0 & 0 & 0 \\ \underbrace{0\cdots0}_{3\,elements} & 0 & I & I & 0 & -2I & 0 \end{pmatrix}\zeta(t) \\ &= \Pi\zeta(t)\end{aligned} \tag{17}$$

Let,

$$\zeta(t) = [e^T(t), G^T(e(t)), \int_{t-\ddot{\gamma}(t)}^{t}\dot{e}^T(s)\,\mathrm{d}s, e^T(t-\ddot{\gamma}_1),$$

$$e^T(t-\ddot{\gamma}(t)), e^T(t-\ddot{\gamma}_2), \frac{1}{\ddot{\gamma}(t)-\ddot{\gamma}_1}\int_{t-\ddot{\gamma}(t)}^{t-\ddot{\gamma}_1} e^T(\omega)\mathrm{d}\omega,$$

$$\frac{1}{\ddot{\gamma}_2-\ddot{\gamma}(t)}\int_{t-\ddot{\gamma}_2}^{t-\ddot{\gamma}(t)} e^T(\omega)\,\mathrm{d}\omega, \int_{t-\ddot{\gamma}_1}^{t} e^T(\theta)\,\mathrm{d}\theta]^T$$

$$M_1 = [\bar{A} + \bar{k} + \sigma\bar{H}\bar{\Upsilon}\bar{W} + \bar{C}\otimes\Gamma + \bar{k}_\tau, \bar{B}, -(k_\tau + \bar{\delta}_1(\bar{C}\otimes\Gamma)), \underbrace{0\cdots0}_{6\,elements}]$$

$$\dot{e}(t) = M_1\zeta(t) + (\sigma(t)-\sigma)\bar{H}\bar{\Upsilon}\bar{W}e(t) + (\bar{\delta}_1 - \bar{\delta}_1(t))(\bar{C}\otimes\Gamma) \times\int_{t-\ddot{\gamma}(t)}^{t}\dot{e}(s)\,\mathrm{d}s$$

From (15), we represent

$$M_2 = \begin{bmatrix} 0 & 0 & (\bar{C}\otimes\Gamma) & 0 & 0 & 0 & \underbrace{0\cdots0}_{3\,elements} \end{bmatrix}$$

$$M_3 = \begin{bmatrix} \bar{H}\bar{\Upsilon}\bar{W} & 0 & 0 & 0 & \underbrace{0\cdots0}_{5\,elements} \end{bmatrix}$$

Hence,

$$\begin{aligned}\mathbf{E}\{L\bar{V}_3(t)\} \le \mathbf{E}\{\zeta^T(t)(M_1^T(\ddot{\gamma}_{12}\hat{S} + \ddot{\gamma}_2Z)M_1 + \bar{\delta}_1(1-\bar{\delta}_1)M_2 \\ \times(\ddot{\gamma}_{12}\hat{S} + \ddot{\gamma}_2Z)M_2^T + \sigma(1-\sigma)M_3^T(\ddot{\gamma}_{12}\hat{S} \\ + \ddot{\gamma}_2Z)M_3) \quad - \frac{1}{\ddot{\gamma}_{12}}\Pi^T\Psi\Pi)\zeta(t)\} \\ - \frac{1-\tilde{\mu}}{\ddot{\gamma}_2}(\int_{t-\ddot{\gamma}(t)}^{t}\dot{e}(s)\mathrm{d}s)^TZ(\int_{t-\ddot{\gamma}(t)}^{t}\dot{e}(s)\,\mathrm{d}s)\}\end{aligned} \tag{18}$$

From some simple computations, Assumption (2.1) can be presented as:

$$-\begin{bmatrix} e(t) \\ G(e(t)) \end{bmatrix}^T\begin{bmatrix} F_1 & -F_2 \\ * & I \end{bmatrix}\begin{bmatrix} e(t) \\ G(e(t)) \end{bmatrix} \ge 0, \tag{19}$$

$$F_1 = I_N\otimes\frac{Z_1^TZ_2 + Z_2^TZ_1}{2}, F_2 = I_N\otimes\frac{Z_1^T + Z_2^T}{2}$$

combining equations (13)-(14), (18)-(19), hence the following:

$$\mathbf{E}\{L\bar{V}(t)\} \le \mathbf{E}\{\zeta^T(t)\Phi\zeta(t)\} \tag{20}$$

This conclude $\mathbf{E}\{L\bar{V}(t)\} < 0$, if $\mathbf{E}\{\zeta^T(t)\Phi\zeta(t)\} < 0$, holds when $\Phi < 0$:
$\Phi = \Omega_1 + M_1^T\Theta M_1 + \bar{\delta}_1(1-\bar{\delta}_1)M_2^T\Theta M_2 + \sigma(1-\sigma)M_3^T\Theta M_3 + \Omega_2 + \Omega_2^T$
From theorem (1), we represent

$$\Omega_2 = \begin{bmatrix} \sigma P\bar{H}\bar{\Upsilon}\bar{W} \\ 0_{8n\times n} \end{bmatrix}\begin{bmatrix} I & \underbrace{0\cdots 0}_{8\,elements} \end{bmatrix}, \Theta = \ddot{\gamma}_{12}\hat{S} + \ddot{\gamma}_2 Z$$

The application of Schur complements formula yeilds the following:

$$\tilde{\Phi} = \begin{bmatrix} \Omega_z & M_1^T P & \Omega_{zz} & \sqrt{\sigma(1-\sigma)}M_3^T P \\ * & -\Theta^{-1} & 0 & 0 \\ * & * & -\Theta^{-1} & 0 \\ * & * & * & -\Theta^{-1} \end{bmatrix} < 0 \tag{21}$$

where, $\Omega_z = \Omega_1 + \Omega_2 + \Omega_2^T$, $\Omega_{zz} = \sqrt{\bar{\delta}_1(1-\bar{\delta}_1)}M_2^T P$. By using the method of congruence transformation with $diag\{\underbrace{I, I, \ldots, I}_{9}, P, P, P\}$ on (21) and utilizing $P\Theta^{-1}P \geq 2P - \Theta$, ensures $\tilde{\Phi} < 0$. Applying simple computations considering Lemma (2.4), let $\Phi_1 = \begin{bmatrix} \Omega_1 & M_{1*}^T P & \Omega_{zz} & 0 \\ * & \Theta - 2P & 0 & 0 \\ * & * & \Theta - 2P & 0 \\ * & * & * & \Theta - 2P \end{bmatrix}$, $M_{1*} = M_1 - \begin{bmatrix} \sigma\bar{H}\bar{\Upsilon}\bar{W} & \underbrace{0\cdots 0}_{8\,elements} \end{bmatrix}$

$N^T = \begin{bmatrix} \sigma(P\bar{H})^T & 0_{n,8n} & \sigma(P\bar{H})^T & 0 & \sqrt{\sigma(1-\sigma)}(P\bar{H})^T \end{bmatrix}$

$W_* = \begin{bmatrix} W & 0_{n,11n} \end{bmatrix}$. Then we have

$$\begin{bmatrix} \Phi_1 & N & \varepsilon W_*^T \\ * & -\varepsilon I & 0 \\ * & * & -\varepsilon I \end{bmatrix} < 0 \tag{22}$$

Based on Lemma (2.4), if the conclusion $\tilde{\Phi} < 0$ is true, then $\mathbf{E}\{LV(t)\} < 0$. Hence, the synchronized error system is asymptotically stable. This completes the proof. □

The following theorem is presented based on the above results in addressing the nonfragile control design problem.

Theorem 2 *Let $\ddot{\gamma}_1, \bar{\delta}_1, \ddot{\gamma}_2, \sigma \in [0,1]$, and $\tilde{\mu} < 1$, be some given scalars. The given CDNs is synchronized, when some symmetric positive definite matrices $P = diag\{P_1, P_2, \ldots, P_N\}, \hat{S} = diag\{\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N\}$ and any matrices $Y_1 = diag\{Y_1^1, Y_2^1, \ldots, Y_N^1\} \in \mathbb{R}^{nN\times nN}$ and $Y_2 = diag\{Y_1^2, Y_2^2, \ldots, Y_N^2\} \in \mathbb{R}^{nN\times nN}$ exit with the positive scalar ε, hence $\bar{\Omega} < 0$, where*

$$\bar{\Omega} \equiv \begin{bmatrix} \bar{\Omega}_1 & \bar{\Omega}_2 \\ * & \bar{\Omega}_3 \end{bmatrix} < 0, \tag{23}$$

$$\bar{\Omega}_1 \equiv \begin{bmatrix} \bar{\Omega}_1^1 & \bar{\Omega}_1^2 \\ * & \bar{\Omega}_1^3 \end{bmatrix},$$

$$\bar{\Omega}_1^1 \equiv \begin{bmatrix} \bar{\Omega}*_1^1 & P\bar{B} + F_2 & -\bar{\delta}_1 P(\bar{C}\otimes\Gamma) - Y_2 \\ * & -I & 0 \\ * & * & -\frac{1-\tilde{\mu}}{\ddot{\gamma}_2}Z \end{bmatrix},$$

$$\bar{\Omega}_1^2 \equiv \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \bar{\Omega}*_1^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \bar{B}^T P & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\bar{\delta}_1(\bar{C}\otimes\Gamma)^T & \wp \end{bmatrix},$$

$$\bar{\Omega}_1^3 \equiv \begin{bmatrix} \bar{\Omega}_1^{31} & \bar{\Omega}_1^{32} \end{bmatrix}, \bar{\Omega}_2 \equiv \begin{bmatrix} 0 & \sigma P\bar{H} & \varepsilon\bar{W}^T \\ 0 & 0 & 0 \\ \bar{K}_\tau^T & 0 & 0 \\ 0_{6n\times n} & 0_{6n\times n} & 0_{6n\times n} \\ 0 & \sigma P\bar{H} & 0 \\ 0 & \tilde{\omega} & 0 \end{bmatrix},$$

$$\bar{\Omega}_3 \equiv \begin{bmatrix} -\Theta^{-}1 & 0 & 0 \\ * & -\varepsilon I & 0 \\ * & * & -\varepsilon I \end{bmatrix}$$

$$\bar{\Omega}_1^{31} \equiv \begin{bmatrix} -4\hat{S} & \Delta & \Delta_1 & 6\hat{S} & \omega_1 \\ * & \Delta_2 & \tilde{\Delta}_1 & \tilde{\Delta}_2 & \tilde{\Delta} \\ * & * & -4\hat{S} & \omega_2 & 6\hat{S} \\ * & * & * & -12\hat{S} & -4L_{22} \\ * & * & * & * & -12\hat{S} \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}$$

$$\bar{\Omega}_1^{32} \equiv \begin{bmatrix} 0_{5n\times n} & 0_{5n\times n} & 0_{5n\times n} \\ -\frac{1}{\ddot{\gamma}_1}W_2 & 0 & 0 \\ * & \omega_3 & 0 \\ * & * & -\Theta^{-1} \end{bmatrix},$$

Where:

$\bar{\Omega}*_1^2 = \bar{A}^T P + Y_1^T + (\bar{C}\otimes\Gamma)^T P + Y_2^T$

$\bar{\Omega}*_1^1 = P\bar{A} + \bar{A}^T P + Y_1 + Y_1^T + P(\bar{C}\otimes\Gamma) + (\bar{C}\otimes\Gamma)^T P + Y_2 + Y_2^T + W_1 + \ddot{\gamma}_1 W_2 - F_1$, $\tilde{\Delta} = 6\hat{S} - 2L_{12} + 2L_{22}$

$\wp = \sqrt{\sigma(1-\sigma)}(\bar{C}\otimes\Gamma)^T$, $\tilde{\Delta}_1 = -2\hat{S} + L_{21} - L_{22} - L_{11} + L_{12}$

$\omega_1 = 2L_{22} + 2L_{12}$, $\omega_2 = 2L_{22}^T - L_{12}^T$, $\omega_3 = \Theta - 2P$

$\tilde{\Delta}_2 = 2L_{12}^T + 6\hat{S} + 2L_{22}^T$, $\tilde{\omega} = \sqrt{\sigma(1-\sigma)}\hat{S}\bar{H}$

$\Delta = -2\hat{S} - L_{11} - L_{12} - L_{21} - L_{22}$; $\Delta_1 = L_{11} - L_{12} + L_{21} - L_{22}$;

$\Delta_2 = -(1-\tilde{\mu})W_1 - F_1 - 8\hat{S} + L_{11} - L_{11}^T + L_{12} - L_{12}^T + L_{21}^T - L_{22}^T - L_{21} - L_{22}$;

Let $Y_1 = P\bar{k}$ and $Y_2 = P\bar{k}_\tau$. The control gains can be derived as $\bar{k} = P^{-1}Y_1$ and $\bar{k}_\tau = P^{-1}Y_2$. The proof is directly obtained from Theorem (1).

4 Numerical Simulations

The objective of this section is to exhibit the correctness of our synchronization schemes in Section (3).

Example 1. The CDNs (1) is considered which comprises of five nodes with each node been two dimensional, such that $N = 5$ and $n = 2$. The other parameters involved are:

$$\mathrm{A} = \begin{bmatrix} 0.7 & 0.2 \\ 0.5 & 1.0 \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix},$$

$$\Gamma = 0.5I_n$$

$$B = \begin{bmatrix} 1.50 & -0.82 \\ 5.15 & 2.50 \end{bmatrix},$$

The controller gain fluctuations made to satisfy $\triangle k_i(t)$ are defined as

$$H_1 = 0.5I_n, H_2 = 0.6I_n, H_3 = 0.7I_n, H_4 = 0.8I_n,$$
$$H_5 = 0.9I_n, W_1 = 0.6I_n, W_2 = 0.5I_n, W_3 = 0.4I_n,$$
$$W_4 = 0.3I_n, W_5 = 0.2I_n$$

Let $f(\tilde{r}_i(t))$ be represented as

$$f(\tilde{r}_i(t)) = \begin{bmatrix} 0.2\tilde{r}_{i1} - \tanh(0.1\tilde{r}_{i1}) \\ 0.1\tilde{r}_{i2} \end{bmatrix}.$$

The above $f(\tilde{r}_i(t))$, satisfies the sector-bounded condition in assumption (2.1) with

$$Z_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \; Z_2 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

Considering this example, we set $\sigma = 0.5$, $\bar{\delta}_1 = 0.5$, $\tilde{\mu} = 0.2$, $\ddot{\gamma}_1 = 0$, $\ddot{\gamma}_2 = 0.5$.
From MATLAB LMI toolbox, Theorem (2) is verified, where feasible solution of the LMIs in Theorem (2) is obtained. The controller gain matrices obtained are:

$$K_1 = \begin{bmatrix} -1.3647 & -0.3460 \\ -0.1196 & -1.5990 \end{bmatrix},$$
$$K_2 = \begin{bmatrix} -1.3121 & -0.3079 \\ -0.0990 & -1.4745 \end{bmatrix},$$
$$K_3 = \begin{bmatrix} -1.2755 & -0.3126 \\ -0.0893 & -1.3825 \end{bmatrix},$$
$$K_4 = \begin{bmatrix} -1.2436 & -0.3139 \\ -0.0836 & -1.3191 \end{bmatrix},$$
$$K_5 = \begin{bmatrix} -1.2226 & -0.3139 \\ -0.0780 & -1.2753 \end{bmatrix},$$
$$K_{1\tau} = \begin{bmatrix} -0.2735 & -0.1907 \\ -0.0188 & -0.1500 \end{bmatrix},$$
$$K_{2\tau} = \begin{bmatrix} -0.3101 & -0.1871 \\ -0.0325 & -0.2592 \end{bmatrix},$$
$$K_{3\tau} = \begin{bmatrix} -0.3163 & -0.1580 \\ -0.0335 & -0.3072 \end{bmatrix},$$
$$K_{4\tau} = \begin{bmatrix} -0.3192 & -0.1247 \\ -0.0292 & -0.3276 \end{bmatrix},$$
$$K_{5\tau} = \begin{bmatrix} -0.3251 & -0.11251 \\ -0.0263 & -0.3376 \end{bmatrix}.$$

Using the following initial conditions, $\tilde{r}_1(0) = [3,-1]^T$, $\tilde{r}_2(0) = [0,1]^T$, $\tilde{r}_3(0) = [-6,2]^T$, $\tilde{r}_4(0) = [3,-2]^T$, $\tilde{r}_5(0) = [-1,1]^T$, and $s(0) = [2,3]^T$. The error trajectories (10) is shown in figure (1) without control input. The control inputs and error system trajectories with control inputs are shown in figures (2) and (3) respectively .

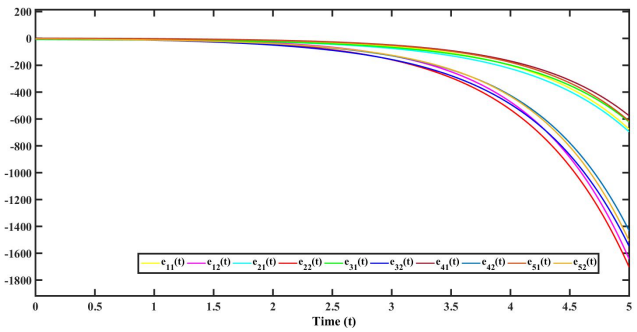


Figure 1: Error synchronization without control inputs $\hat{u}(t)$ in Example 1.

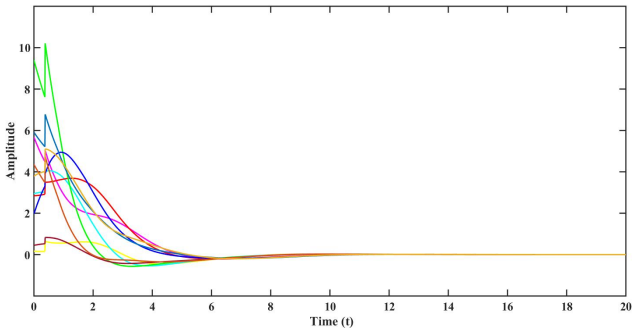


Figure 2: Control inputs $\hat{u}(t)$ in Example 1.

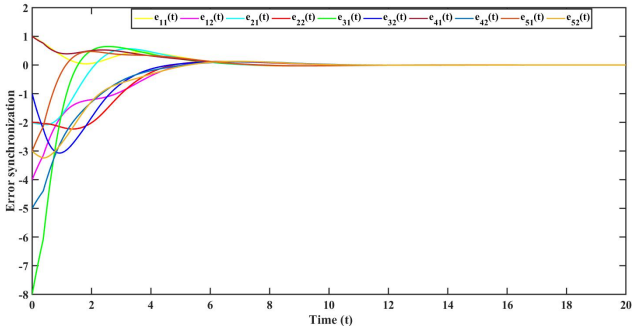


Figure 3: Error synchronization with control inputs $\hat{u}(t)$ in Example 1.

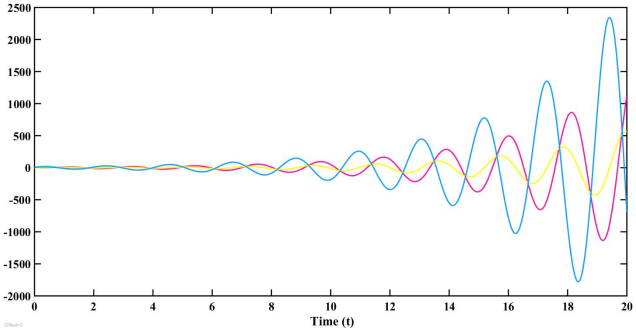


Figure 4: Error synchronization without control inputs $\hat{u}(t)$ in Example 2.

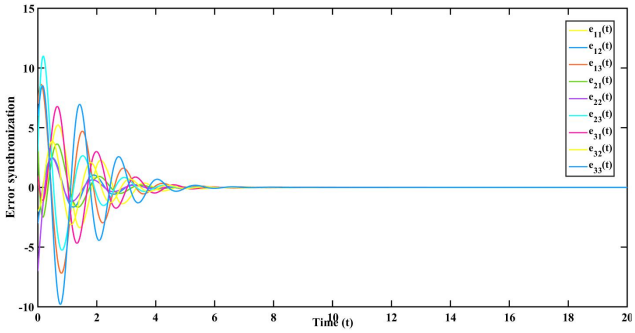


Figure 5: Error synchronization with control inputs $\hat{u}(t)$ in Example 2.

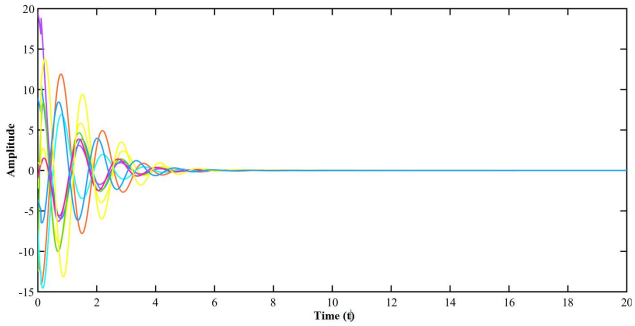


Figure 6: Control inputs $\hat{u}(t)$ in Example 2.

Example 2. Chua's circuit is adopted in this example as an isolated node which is decribed by the following equation:

$$\begin{cases} \dot{s}_1 = & \bar{\kappa}(s_2 - s_1 + \bar{\varphi}(s_1)) \\ \dot{s}_2 = & s_1 - s_2 + s_3 \\ \dot{s}_3 = & -\bar{b}s_2. \end{cases} \tag{24}$$

Let $A = 0$, $B = I_3$, $\delta_1 = 0.1$, $\bar{\kappa}$ =10, $\bar{b}$ =14.87 and $\bar{\varphi}(s_1) = \tilde{\omega}_1 s_1 + \frac{1}{2}(\tilde{\omega}_2 - \tilde{\omega}_1)\bar{\psi}(s_1)$ where $\tilde{\omega}_1 = -0.68$, $\tilde{\omega}_2 = -1.27$, and $\bar{\psi}(s_1) = (|s_1 + 1| - |s_1 - 1|)$. Denote $s = [s_1, s_2, s_3]^T$, $\bar{\phi} = -\frac{1}{2}(\tilde{\omega}_2 - \tilde{\omega}_1)$, where

$$f(s) = \begin{bmatrix} -\bar{\kappa}(1-\tilde{\omega}_1) & \bar{\kappa} & 0 \\ 1 & -1 & 1 \\ 0 & -\bar{b} & 0 \end{bmatrix} + \begin{bmatrix} \bar{\phi}\bar{\psi}(s_1) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

From Assumption(2.1),

$$Z_1 = \begin{bmatrix} 2.7 & 10 & 0 \\ 1 & -1 & 1 \\ 0 & -14.87 & 0 \end{bmatrix}, Z_2 = \begin{bmatrix} -3.2 & 10 & 0 \\ 1 & -1 & 1 \\ 0 & -14.87 & 0 \end{bmatrix}.$$

The inner coupling Γ and network topology $\bar{C}$ matrices are given as

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \bar{C} = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & -2 \end{bmatrix}.$$

The parameters considered are: $\ddot{\gamma}_1 = 0.01, \ddot{\gamma}(t) = 0.4 + 0.01\sin(10t), \ddot{\gamma}_2 = 0.41, \tilde{\mu} = 0.1, \sigma = 0$. MATLAB LMI toolbox is used on Theorem (1). The obtained controller gain matrices are

$$k_1 = \begin{bmatrix} -1.6520 & 0.3835 & 0.0811 \\ 0.6663 & -0.9311 & 0.7839 \\ 0.0602 & 0.43138 & -1.3760 \end{bmatrix},$$

$$k_2 = \begin{bmatrix} -2.6121 & 0.3850 & 0.2110 \\ 0.5820 & -1.9887 & 1.1084 \\ 0.0163 & 0.5489 & -1.1378 \end{bmatrix},$$

$$k_3 = \begin{bmatrix} -0.7934 & 0.3540 & 0.0905 \\ 0.7406 & 0.6739 & 0.2685 \\ 0.0820 & 0.2063 & -0.5514 \end{bmatrix},$$

$$k_{\tau 1} = \begin{bmatrix} -0.1015 & 0.0165 & -0.0144 \\ -0.0051 & -0.5132 & 0.3061 \\ -0.0022 & 0.1375 & -0.1897 \end{bmatrix},$$

$$k_{\tau 2} = \begin{bmatrix} -0.0967 & -0.0076 & -0.0082 \\ -0.0175 & -0.3802 & -0.0082 \\ -0.0007 & -0.0081 & -0.1502 \end{bmatrix},$$

$$k_{\tau 3} = \begin{bmatrix} -0.0810 & 0.0670 & -0.0145 \\ -0.0070 & -1.2975 & 1.2294 \\ -0.0177 & 0.5618 & -0.0825 \end{bmatrix}.$$

Assume the following initial conditions are considered for the system: $s(0) = [0, -1, 1]^T$ $\tilde{r}_1(0) = [1, -2, 8]^T, \tilde{r}_2(0) = [4, -6, 4]^T$, *and* $\tilde{r}_3(0) = [1, -1, 7]^T$. The simulation result given in figure (4) indicates state error trajectories without control input, whereas that of figures (5) and (6) depict the synchronized closed-loop error system and the control input signals respectively.

5 Conclusion

This paper shows how an appropriate Lyapunov Krasovskii functional is used to address a non fragile synchronization control problem for a stochastic coupling complex dynamical networks with time-varying delays and control gain perturbations. The application of extended version of Jensen's inequality ensured the LMIs to be feasible. Finally, numerical examples with simulations are shown to illustrate the validity and applicability of our proposed control scheme.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to express their profound gratitude to Professor Zhong Shouming of UESTC for his immense support and contributions.

References

[1] F. A. Gyamfi, Y. Cheng, C. Yin and S. Zhong, Synchronization of complex dynamical networks with randomly coupling via nonfragile control, 2018 Tenth International Conference on Ubiquitonus and Future Networks (ICUFN), Prague, 2018, pp. 497 - 502.

[2] D. J. Watts and S. H. Strogatz,Collective dynamics of small-world networks.Nature,393(6684)(1998) 440 - 442.

[3] W. Yu, G. Chen, J. Lü, On pinning synchronization of complex dynamical networks , Automatica 45 (2) (2009) 429 - 435.

[4] R. Rakkiyappan, G. Velmurugan, J. Nicholas George, R. Selvamani, Exponential synchronization of Lur'e complex dynamical networks with uncertain inner coupling and pinning impulsive control, Applied maths. and Comput., 307(2017) 217 - 231.

[5] J. Wang, H. Zhang, Z. Wang and B. Wang, Local exponential synchronization in complex dynamical networks with time-varying and hybrid coupling,Applied Mathematics and Computation,225(2013) 16 - 32.

[6] D. Li, Z. Wang and G. Ma, Controlled synchronization for complex dynamical networks with random delayed information exchanges:A non-fragile approach. Neurocomputing,171(2016) 1047 - 1052.

[7] M. Fang, Synchronization for complex dynamical networks with time delay and discrete-time information. Applied Mathematics and Computation,258(2015) 1 - 11.

[8] M. J. Park, O. M. Kwon, J. H. Park, S. M. Lee and E. J. Cha, On synchronization criterion for coupled discrete-time neural networks with interval time-varying delays. Neurocomputing, 99(2013) 188 - 196.

[9] Y. Wang, Y. Ma, A. Chen, Exponential synchronization of Markovian jump complex dynamical networks with partially uncertain transition rates and stochastic disturbances. Neurocomputing 304 (2018) 30 - 46.

[10] Y. Tang, F. Qian, H. Gao and J. Kurths, Synchronization in complex networks and its application- A survey of recent advances and challenges. Annual reviews in control 38(2) (2014) 184 - 198.

[11] A. Arenas, A. Diaz - Guilera, J. Kurths, Y. Moreno and C. Zhou, Synchronization in complex networks, Physics Reports 469 (2008) 93 - 153.

[12] M. Xu, J - L. Wang, Y - L. Huang, P - C. Wei and S - X. Wang, Pinning sychronization of complex dynamical networks with and without time-varying delay. Neurocomputing, 266(2017) 263 - 273.

[13] F. Adu - Gyamfi, Y. Cheng, C. Yin and S. Zhong, Exponential H_∞ synchronization of non-fragile sampled-data controlled complex dynamical networks with random coupling and time varying delay. System Science and Control Engineering, 6(1) (2018), 370 - 387.

[14] M. Syed Ali, J. Yogambigai and J. Cao, Synchronization of master-slave markovian switching complex dynamical networks with time-varying delays in nonlinear function via sliding mode control. Acta Mathematica Scientia, 37(2)(2017) 368 - 384.

[15] M. Yang, Y -W. Wang, J - W. Xiao and Y. Huang, Robust synchronization of singular complex switched networks with parameteric uncertainities and unknown coupling topologies via impulsive control. Communications in Nonlinear Science and Numerical Simulation, 17(11) (2012), 4404 - 4416.

[16] V. M. Alfaro, R. Vilanova and O. Arrieta, Fragility analysis of PID controllers. IEEE Control Applications, (CCA) Intelligent Control, (ISIC), (2009) 725-730.

[17] Y. Liu, B - Z. Guo, J. .H. Park and S - M. Lee, Nonfragile exponential synchronization of delayed complex dynamical networks with memory sampled-data control. IEEE Transactions on Neural Networks and Learning Systems. 29(1) (2018) 118 - 128.

[18] R. Zhang, D. Zeng, J. .H. Park, Y. Liu and S. Zhong, Nonfragile sampled-data synchronization for delayed complex dynamical networks with randomly occurring controller gain fluctuation. IEEE Transactions on Systems, Man and Cybernetics: Systems, 48(12)(2018) 2271 - 2281.

[19] E. Gyurkovics, K. Kiss and A. Kazemy, Non-fragile exponential synchronization of delayed complex dynamical networks with transmission delay via sampled-data control. Journal of the Franklin Institute, 355(17)(2018), 8934 - 8956.

[20] V. M. Revathi, P. Balasubramaniam and K. Ratnavelu, Delay-dependent H 1 filtering for complex dynamical networks with time-varying delays in nonlinear function and network couplings. Signal Processing,118(2016) 122 - 132.

[21] X. Li and X. Fu, Effect of leakage time varying delay on stability of nonlinear differential systems. Journal of the Franklin institute, 350(6)(2012) 1335 - 1344.

[22] H. Zhang, J. Wang, Z. Wang and H. Liang, Mode-Dependent stochastic synchronization for Markovian coupled neural networks with time-varying mode-delays. IEEE Transactions on Neural Networks and Learning systems,26(11)(2015) 2621 - 2634.

[23] H. Bao, J .H. Park and J. Cao, Exponential synchronization of coupled stochastic memristor-based neural networks with time-varying probabilitic delay coupling and impulsive delay. IEEE Transactions on Neural Networks and Learning Systems,27(1)(2015) 190 - 201.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 17-26 (2019)

www.astesj.com

Special Issue on Advancement in Engineering and Computer Science

ASTESJ

ISSN: 2415-6698

A Wearable Exoskeleton Rehabilitation Device for Paralysis – A Comprehensive Study

Ahmed Roshdy, Samer Al Kork*, Sherif Said, Taha. Beyrouthy

College of Engineering and Technology, American University of the Middle East, Kuwait

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 27 December, 2018
Online : 20 January, 2019

Keywords:
Rehabilitation
EEG
Prosthetic
Emotiv headset
Paralysis

ABSTRACT

As the technology grows scientists and engineers are trying to combine their work to compensate some body parts that is lost. Prosthetic devices grabbed the attention of most of the doctors and engineers working on solution for lost body parts. Generally prosthetic devices are either external wearable devices or internal ones. Such devices may depend on a built on microcontroller or the brain signals from the patient himself. Although it is used for lost body parts it can also be used for rehabilitation, power assistance, diagnostics, monitoring, ergonomics, etc. The currently used devices usually have the disadvantages of big size and high cost. The suggested device in this paper is targeting the design of portable rehabilitation device with light weight and low cost for paralyzed hand people. The suggested device allows the user to train the patient's hand or perform some needed exercises for his impaired hand. This helps the user to restore the normal hand movement and functionality. The device includes two modes of operation to be chosen by the user through the platform built on a microprocessor which can help controlling the exoskeleton to perform the needed exercises or tasks. Collaboration with several healthcare organizations will be considered to verify or test the effectiveness of this exoskeleton.

1. Introduction

This paper is an extension of work originally presented in BioSMART, the 2nd International Conference on Bio-engineering for Smart Technologies" titled 'A wearable rehabilitation device for paralysis' [1]. Diseases like stroke may lead to the loss of some parts of the human body and some other people are actually suffering of paralysis, broken bones, spinal cord injury, hemiplegia or traumatic brain injury. People facing such diseases or injuries are actually facing some problems in their day to day activities because of the infected or injured body part. One of the suggested solutions for such cases is the rehabilitation therapy to improve the infected part movement and regain the strength and power of that part. Some of the physiotherapy exercises are needed for the recovering rehabilitation process for the patient to be able to completely or partially restore the normal movement of the paralyzed body part [2]. Although this process is needed for hands or legs rehabilitation, this paper is actually focusing on a device to be used for the hands' problems and injuries. Humans are using their hands for writing, touching, holding, squeezing, and so many other functionalities. The partial or total loss of this part of the human body means losing so many valuable functions that any human being needs. The full hand with all its many joints act as natural motors to perform routine daily tasks [2]. This paper is meant to target a rehabilitation therapy system for patients with hand disabilities like too weak hands or even paralysis. The suggested device compared to the existing devices is considered to be low cost, lightweight, customizable, programmable and safe for human practice. All the functions of the device should be running in two different rehabilitation modes of training called active and passive. The design process will move into four phases, which are human interface, processing unit, health care sensors, and display. Each one of them has special task to do. Human interface will be responsible of the interface between the user and the processing unit. Processing unit will transmit commands to the electromechanical part. Electromechanical part will control the movement of the Exoskeleton. Health 10 care sensor will be used to monitor the user's health to insure the safety. The results will be shown in the display part.

The paper is going through literature review for common wearable devices available in the market showing the mode of control and the controlling method of each one. Also the literature review is covering the different brain waves' readers as well as the brain waves and the motor controls on the device itself. At the end the proposed wearable device is discussed mainly in hardware with a comparison to the available devices.

2. Literature Review

2.1. Wearable Device

Here in this section is the comparison between the main projects already exist as rehabilitation wearable devices and assistive exoskeletons. In Table 1 the comparison shows the difference in the modes and the methods used in some projects

*Samer Al Kork, American University of the Middle East, +965 2225 1400 Ext. 1732, samer.alkork@aum.edu.kw

that were built for rehabilitation and assistive purposes for either hand or leg usage. Most of the projects are based in using either Virtual Reality (VR) or electromyography (EMG) as an interface for the device. None of the currently used devices has built the system using Electroencephalography (EEG) interface for the wearable device like what is offered in this study. The main problems for the devices already exist in the market are the high cost and the heavy weight because of the heavy materials used in their fabrication [4]. Although there already exist so many devices but only few of them offers the two modes of control passive and active. Herewith the study is based on EEG interface with the two possible modes of operation as well as keeping a light affordable weight and low cost.

Table 1 Some Applications of Wearable Devices

Title	Control Modes	Methods
Design and development of a hand exoskeleton for rehabilitation following stroke [4]	Active	-
Hand Rehabilitation Support system Based on Self-motion control, with Clinical case report **Error! Reference source not found.**	Active	VR
Current Hand Exoskeleton Technologies for Rehabilitation and Assistive Engineering **Error! Reference source not found.**	Passive	VR
Design and Development of a Hand Exoskeleton Robot for Active and Passive Rehabilitation [7]	Active/Passive	EMG
An EMG-driven exoskeleton hand robotic training device on chronic stroke [8]	Active	EMG
A review of technological and clinical aspects of robot-aided rehabilitation of upper-extremity after stroke [9]	Active	VR

Hong Kong polytechnic University designed a portable set of robotic hand exoskeleton that can be used or carried anywhere to help stroke patients in opening or closing their hands [8]. It works in an active mode and has 2 degree of freedom for each finger at the MCP and PIP. The robotic hand consists of Velcro straps to hold the hand in place, 5 linear actuators for fingers and a palm support platform. Finger assemblies are used to provide finger's flexion and extension. Also, it is designed in a way that can be used for different finger length. Also, there is an embedded controller that handles the robotic hand various tasks and monitor the EMG signals (sensor) that are used for the closing and opening process of a hand. Moreover, the set contains a wireless remote control system that helps the therapist to configure and select from different training modes (Figure 1).

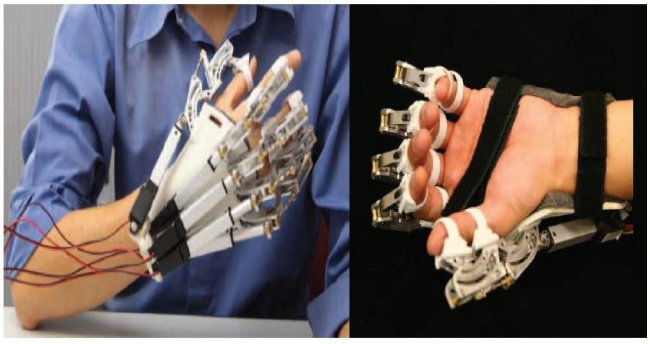


Figure 1. The Set of Hand Exoskeleton and Overview of the Prototype

University of Technology in Sydney Designed and developed a hand exoskeleton for rehabilitation following stroke [4]. The device is achieving full flexion/extension motion of the five fingers of the left hand (impaired hand) based on the motion of the identical digits of the right hand (healthy hand). The device is less than 2 Kg and they chose Aluminum because of its lightweight. The hand exoskeleton has 15 degrees of freedom (DOFs). However, the hand exoskeleton cannot perform abduction/adduction movement; as a consequence, more work needs to be done on the device (Figure 2).

Figure 2. Hand Exoskeleton for Rehabilitation

Kyushu University proposed a hand exoskeleton using three-layered sliding spring mechanism [10]. The idea of this paper is to present a lightweight and compact device to easily use, so under-actuated mechanism is used to reduce the mass and size of the device by limiting the number of actuators into one (Figure 3).

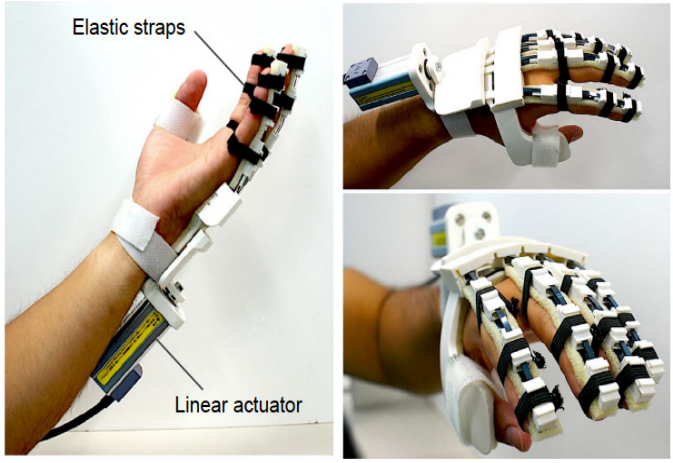


Figure 3. Overview of the Hand Exoskeleton

The weight of the exoskeleton is 320 g. Each finger has three DOF, which is flexion/extension, but the thumb is fixed for the sake of robust grasping. Therefore, the three DOF is actuated through one actuator.as a result; the four fingers will work simultaneously. Three-layered sliding spring consists of 3 springs that is divided into inner (Si), center (Sc) and outer (So) springs and rigid bodies into tip (Rt), inner (ri) and outer (Ro) parts. The mechanism helps to perform flexion motion (Figure 4).

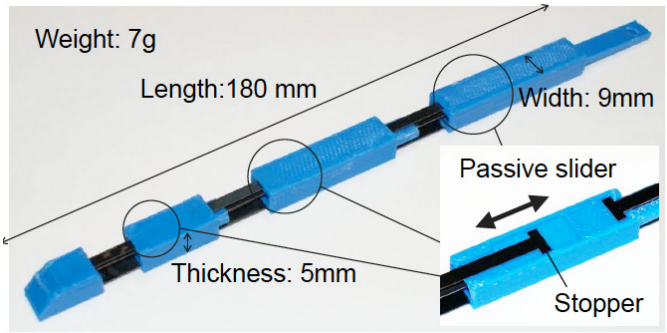


Figure 4. Three-Layered Sliding Spring Mechanism

University of Salford developed a hand assistive exoskeleton that operates in the active mode [11]. It uses virtual reality exerciser to perform the physical therapy exercises. It allows the patient to do the therapy exercises through fun interactive games. This exoskeleton allows hand motion, analysis and recording easily. The exoskeleton has 7 active DOF (Figure 5).

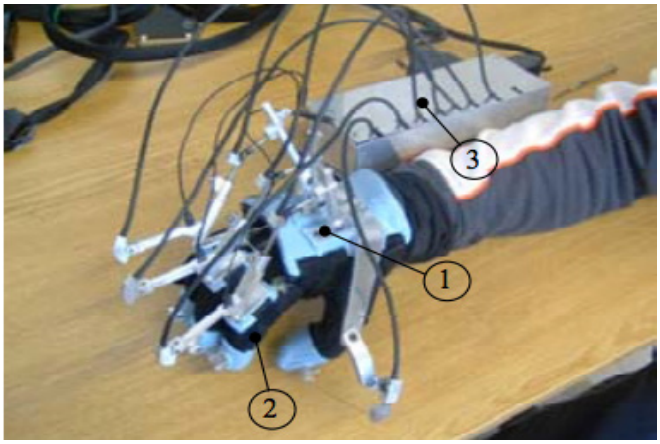


Figure 5. Assistive Exoskeleton

A team at Twenty university develop a low cost portable hand exoskeleton for assistive and rehabilitation in 2016 (Figure 6). They developed it to be for active rehabilitation, cheap, wearable, and portable [12]. This hand controlled by the mussels, they used EMG method. It can be for different sizes. They used 3D printer for some damages exoskeleton's components. The aim of this robotic hand exoskeleton is to assist persons with hand opening disabilities.

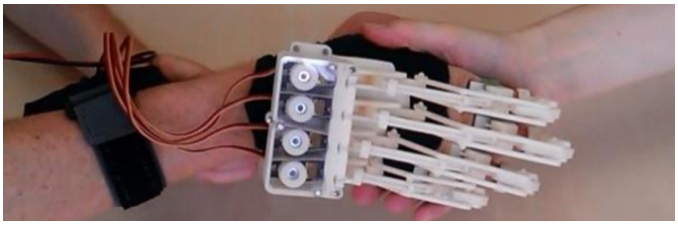


Figure 6. Assistive/Rehabilitation Exoskeleton

2.2. EEG

EEG stands for "Electroencephalography". EEG is a technique used to screen and record electrical signals produced inside the brain (Figure 7). There are two EEG techniques, which are non-invasive and invasive. In the non-invasive technique, electrodes are utilized to record the signals, and they are attached directly to the human head. On the other hand, the invasive technique records the signals by embedding electrodes inside the skull itself till the brain. In addition, there are two different types of communication, the first one is the BCI (brain computer interface), and the second one is BMI (brain machine interface). These two types define the communication method or interface between the brain and the device meant to record the brain signals. Relating the mentioned two communication methods to EEG techniques the BCI communication type follows the non-invasive technique while the BMI type follows the invasive method. Using the EEG will allow the user to control the motors by his own brain signals. The brain signal captured by the EEG is to be sent to set of motors to perform the needed movement or action. These days, this technique has turned out to be affordable and available to general society. Many different products already available in the market are using EEG in order to monitor the brain activity.

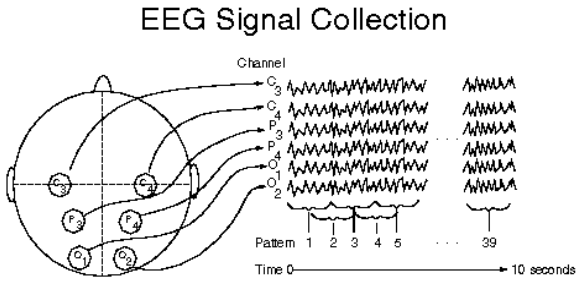


Figure. 7. EEG electrodes record signals from brain

The brain is a very complex and essential organ that its main function is to control the whole body. In general, the brain anatomy or structure is consisting of three main sections these are forebrain, midbrain and hindbrain. The forebrain is classified into three parts, which are thalamus, cerebrum, and hypothalamus. The cerebrum represents the largest part of the brain. It is divided into 4 portions called lobes, which are frontal (F), parietal (P), temporal (T), and occipital (O) lobes.

Each area or region has its own function. The frontal lobe is responsible for problem solving, speaking, judgment, emotional expression, thinking, planning and movement.
Parietal lobe is more into processing sensory and interpreting visual information (reacting into environment); it allows sensation from muscle and skin, body orientation, and reading. Temporal lobe is associated with behavior, memory, hearing and understanding language. Moreover, the occipital lobe is used for color perception, image recognition and vision or sight (Figure 8).

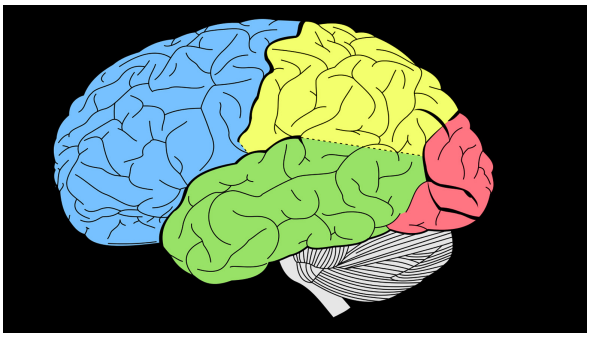


Figure 7. The Main Portions of the Brain

2.3. Emotiv (EPOC+) and the 10-20 System

EPOC follows the 10-20 system electrode placements which is considered to be an international system [13]. This system depends on 10% or 20% separation between the electrodes. It permits 21 electrodes on the surface of the scalp shown in Figure.. Each possible location for the electrodes is described by a letter to identify the lobe and a number to define the location of the hemisphere (left or right side of the brain). It shows the 14 channels of EPOC are distributed as frontal (AF3, AF4, F7, F3, F4, F4), front-central (FC5, FC6), parietal (P8, P7), temporal (T8, T7) and occipital (O1, O2) lobes [14]. The odd numbers refer to the left side of the brain and the even numbers refer to the right side of the brain. The system renamed some locations or points, which are T4, T3, T5 and T6 to be T8, T7, P7 and P8 respectively. Also, the AF and FC are intermediate sites; AF is between Fp and F but FC is between F and C. Moreover, A1 and A2 are the locations of earlobes, DRL and CMS correspond to P3 and P4, which are the reference sensors (Figure 9).

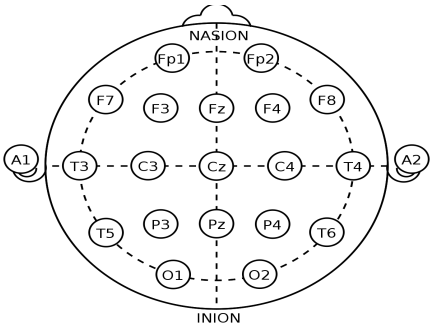


Figure.9. Locations of the reference sensors

2.4. The Motor Area

As the different brain parts are generating different types of waves for each task needs to be performed by the human body, the headset. The primary motor cortex, which is located in the frontal lobe, is responsible for controlling the execution of movement [15]. In fact, it is located in an area called Precentral gyrus. This part of the brain is participating in controlling the movement of different parts of the body like arm, hand, face, foot, etc. (Figure 10). According to the 10-20 system brain map, C3, Cz and C4 are the nearest to the location where motor execution occurs [15], [16]. Unfortunately, the three motor locations aren't available in the EPOC headset. However, there is an easy way to obtain one of the three points; the headset can be tilted a little bit till pointing one of the electrodes to one of the needed three locations.

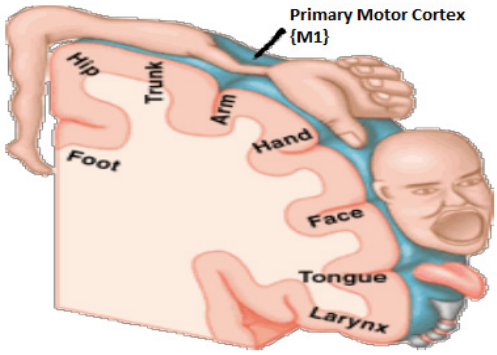


Figure 10. The Location of the Hand Movement in the Motor Cortex

2.5. Brainwaves

The brain signals are different according to the band of frequencies of each one of them. The different frequencies or to be more precise the different band of frequencies are covering all the brain activities whether the human body is moving or not. Each wave generated by the brain is does have particular location to be produced from. The main 5 range of frequencies from the brain are Delta, Beta, Theta, Alpha, and Gamma (Figure 11). Each band has a specific frequency range but a different meaning.

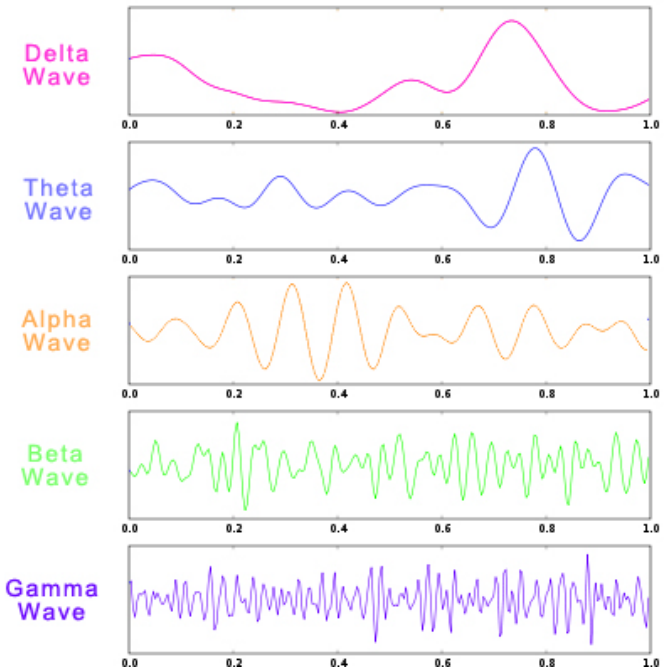


Figure 11. The Main 5 Brainwaves

The first range of frequencies is from 0.5 to 3 Hz and considered to be the Delta range. This range of waves are created while sleeping or relaxing meditation sessions. Also in our scope of work the Delta waves are generated during the normal human body healing during a deep sleep. The next range of waves is the Theta waves and it's in the range of 4 – 8 Hz. Usually these kind of waves generated while sleeping as well or laziness. Third range for the Alpha waves is from 4 to 12 Hz and the got generated by the occipital, parietal and frontal lobes in some cases like closing the eyes and by the frontal lobe during relaxation. Out of the Alpha waves range there is a range from 8-12 Hz is called Mu waves. The Mu waves mainly corresponds to the body movement. In other words, the Mu waves are generated when the human body is relaxed. The next range from 12 to 25 Hz is called Beta waves. Beta waves are generally created in body movements and brain activities like thinking or problem solving. Above 25 Hz signals are called Gamma signals and they are faster than the other waves as noticed from its frequency. Such signals are usually generated with excessive brain work during multi-tasking.

2.6. Virtual Reality

Virtual reality represents an alternative way to be engaged or involved in the rehabilitation therapy routines and not easily get bored by using the rehabilitation gaming system [18],[19]. The rehabilitation gaming system is a virtual reality tool in a three-

dimensional world that let the patient does the therapy exercises through fun interactive games (Figure 12). Also, it can be used as a stimulus to help the patient in the motor imagery process. The VR technology is currently a very up-to-date trend. It is quickly moving towards the mainstream that even big companies are investing in this technology like Samsung, Sony, Apple, etc. They are developing their existing devices to follow this new technology.

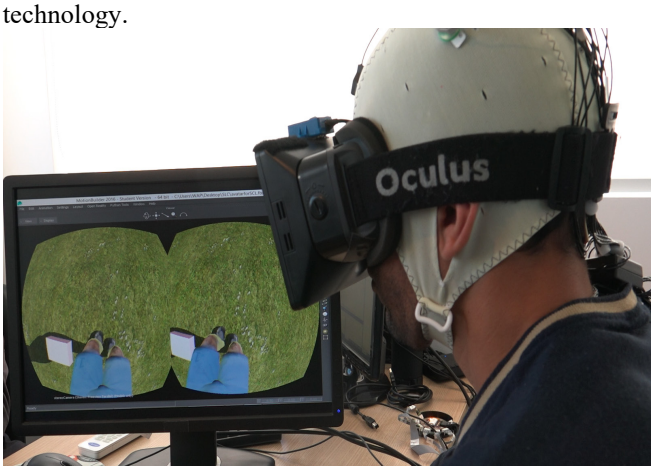


Figure 12. Virtual Reality Tool in a Three-Dimensional World

2.7. Electromyography (EMG)

Electromyography EMG measures electrical current inside muscles (Figure 13). EMG is controlled by nervous system. It gains noise while moving through different tissues. Specific Algorithms should be used to reduce the noise in order to obtain accurate EMG signal [20]. Many applications have been implemented using EMG method for exoskeleton hand control; they used muscle signals to control the impaired hand which is driven by an intact organ [8].

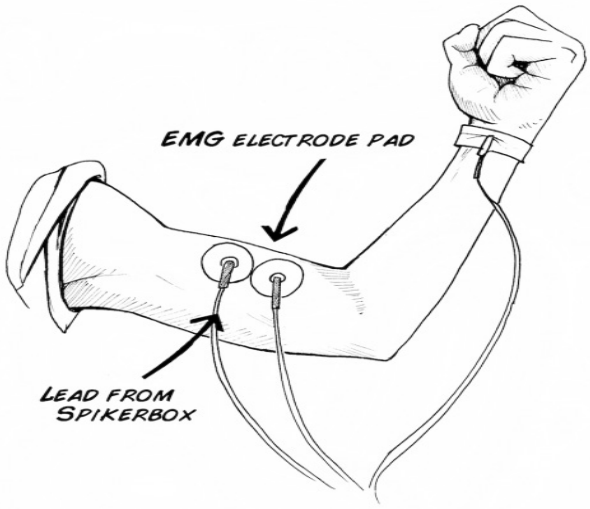


Figure 13. EMG Signals Comes From Muscles

3. Detailed design

3.1 Design Overview

The whole design is described in Figure 18 showing the whole design different stages through the design. In the active mode the EEG headset starts to read the brain signals and detects the location of the signal and decide the lobe and the side of the brain sending the signal. Next step the headset will wirelessly communicate with the computer through Wi-Fi to start the analysis. After the analysis is done the signal should be compared to previously recorded data to indicate the meaning of the brain signal based on pattern recognition. In the Preprocessing unit, the signal will be filtered. Then the signal goes to the second part where features and specific characteristic will be detected. The final stage the signal will go through is the classification where the imagined movement will be identified and performed. After signal processing, the information taken from the headset will be compared to the previous saved data. Using pattern recognition if the comparison got a match, then it will send the data through Bluetooth ZigBee module to the electromechanical parts to start the movement. This part will then give the instructions to the servo motors to start performing the needed movement. This part includes two big servos that ensure the movement of the arm up and down (Figure 14).

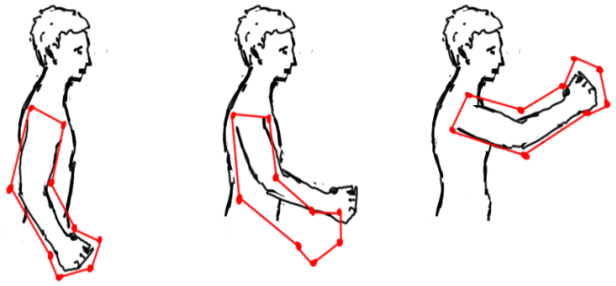


Figure 14. The movement of the arm

The health sensors part includes a body temperature sensor, which allows the user to measure his body temperature [20]. It includes also a pulse oximetry sensor that indicates the arterial oxygen saturation of functional hemoglobin. An LCD screen is present to show the results of some sensors like the body temperature sensor and it is considered to be a major part of the interface between the user and the device. Wi-Fi is used to send all the results to the mobile application. These results will lead to activating or deactivating the mechanical parts (motors). The schematic of the Robotic-Based Rehabilitation system with all its basic parts is shown in Figure 15.

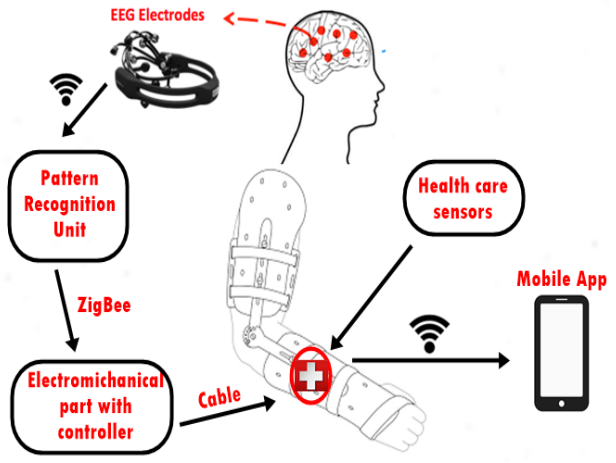


Figure 15. Design overview

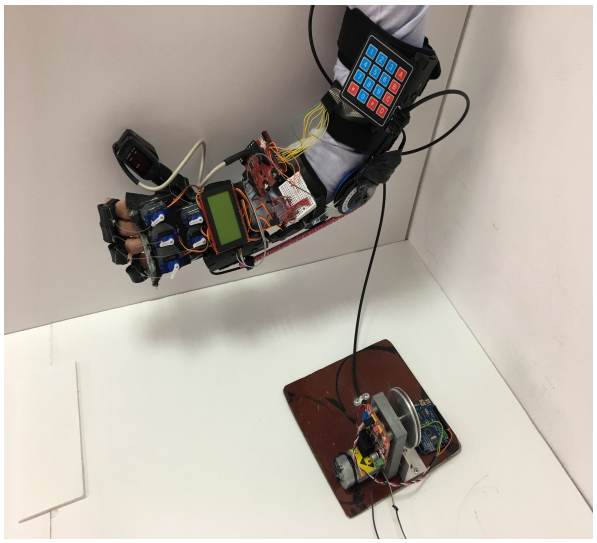


Figure 16. Design schematic

3.2 System Design

The overall design and flowchart is represented in Figure 16 and Figure 17. The user enters his/her age. Then, the temperature, ECG, and Pulses of the user are displayed on the LCD. If the values are in the normal ranges, the system continues running, otherwise, it stops. There are specific ranges for each one of them. The normal body temperature tends to be between the range 36.1°C to 37.2°C. On the other hand, the normal human pulses can be classified according the ages to be 40 to 60 pulses for ages between 1 to 10 years and 60 to 100 pulses for ages between 10 to 60 years. Therefore, these things will provide extra security for the user.

Second, the user picks the required mode through the built on switch (left for active, middle for off, right for passive). If the user selected the first mode which is the passive mode, the user will be able to select the needed exercise through the keypad connected to the Arduino microcontroller. After that, the servo motors will start performing the corresponding movements based on the code on the Arduino board. The motors are connected through rods to the wearable device parts like fingers and rest to perform the right movement.

The other mode of operation is the Active mode that can be selected by moving the switch to the left. In the Active mode the movements and therapy will be controlled by the brain signals read by the EEG headset. When the headset reads the brain signal it will pass it to the Arduino for analyzing to recognize the required move Figure 18. In the beginning the signals are filtered whenever read by the EEG to get rid of the noise signals through bandpass filter and notch filter used particularly for 50Hz signals that will be there most of the time from the surroundings and power sources around the patient [14]. After having the clean signal, the next step is getting the information from the signal and translate it into movements or exercise through the motors connected to the wearable device. The signals movement translated orders are saved in a features vector to be performed. The extracted information is classified in order to help in extracting the features into spatial and spectral information. The location of the electrode extracting the signal is considered to be spectral type of information. On the other hand, spectral information represents the power of the frequency bands. At the end a computer will classify the signals to be able to select the movement. There are many classifiers; however, neural network and SVM are most commonly used [23][24]],[14],[26]. The classifier will then save the obtained data in the feature vector with either movement to class A or class B. Comparing these data with the database stored from pre imagined movements will lead in case of matching to moving the signal to Arduino as the brain for the servo motors to perform the exercise and move the device.

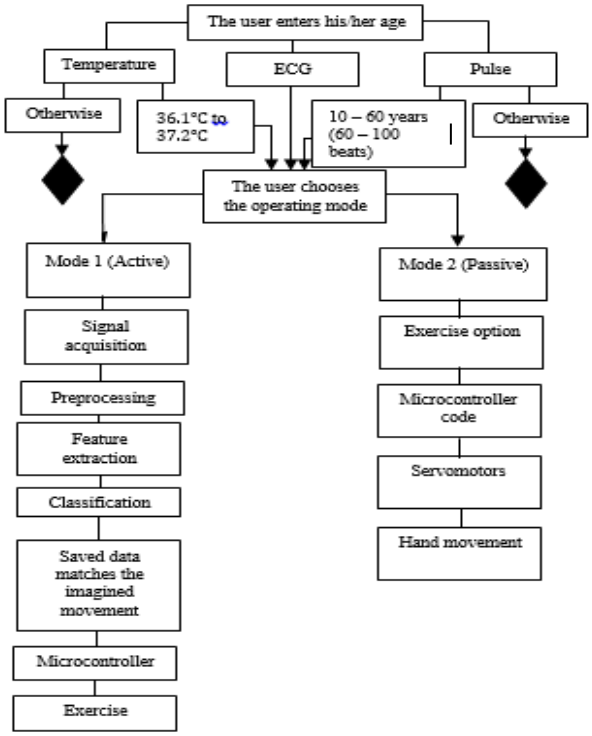


Figure 17. Flowchart of the system design

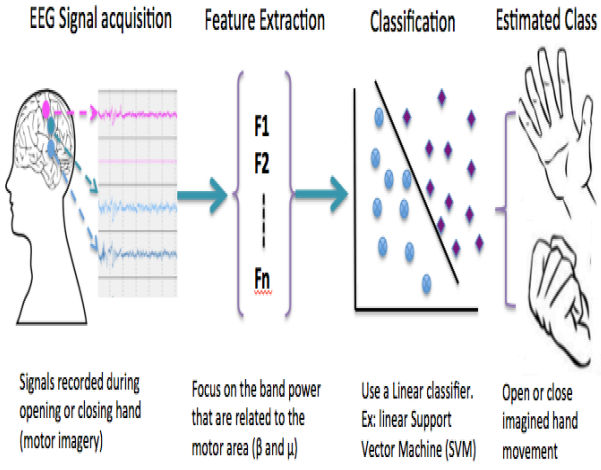


Figure 18. Signal processing steps [21][24]

3.3 Relevant Engineering Application and Calculations

Servomotors was chosen for this project. Four servo motors in two separate joints are needed for each arm. To check the loads

on the motors and make sure about the safety of the human body, it was designed to allow the device to hold weights up to 10 kg. This amount of weight causes the torque on the servo motors while moving. The angle of movement should also be taken into considerations to have the right calculations of the motors mechanical loads. The torque on the servomotors according to the applied force will be calculated using equation (1)

$$\tau = F \times r \times sin\emptyset \quad (1)$$

Where τ is the applied torque, F is the force cause by the weight carried ($F = m \times g$), r is the distance, and $\emptyset$ is the maximum angle of movement which is 90°. The distance used in our design is 0.74m and the mass of the arm is 3.628kg, and the gravity in Kuwait is 9.793N/A. The resulting torque is 26.29N.m. For such torque and safety, the servomotors will be chosen accordingly. From equation (1) there is a direct proportional relation between the torque and the applied force. Taking into considerations the power consumption, the no load power can be calculated using equation 2 knowing that the maximum current used with no load is 500Ma.

$$P = V \times I = 25 \times 0.5 = 12w \quad (2)$$

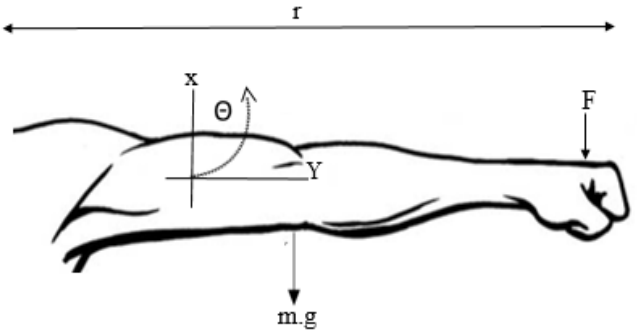


Figure 19. Power Consumption Parameters

3.4 Materials

For the process of building the hand exoskeleton device, so many materials need to be studied in terms of weight, safety, and flexibility in order to choose the suitable one for the project. Iron, aluminum and 3D printed material which is mainly plastic-based are different materials that can be used for exoskeleton. Each one of them has different properties. Iron is heavier than aluminum because of its density, which is 7.1 g/cc. Moreover, Iron is cheaper than aluminum because it does not need many processes to obtain as the aluminum. It is stronger than aluminum. Iron's strength is 169 (MPa) whereas aluminum is 80 (MPa). Aluminum takes time to be manufactured. The features of 3D printer materials are: cheap, flexible, durable, lightweight and available with varies colors. Easy to redesign and model if needs to do any changes. For example; if something is broken we can have reprinted it easily. Furthermore, it is not taking much time to print. The materials that should be used must be environment friendly to provide safety for a long period time. Aluminum is not environment friendly. The releasing of perfluorocarbons during the aluminum smelting process are 9,200 times more harmful than carbon dioxide. Iron is not recommended due to the chemicals which enters into the environment and affect the eco balance, and cause so many health problems. 3D printer materials like Acrylonitrile Butadiene Styrene (ABS) and Polylatic Acid (PLA) can be considered as environment friendly materials because it is easy to recycle, no waste of materials, and no more unsold product. The comparison is summarized in Table 2 &

Table **3**

Table 2. Materials' Specifications

Materials	Price per kg ($)	Strength (Mpa)	Density	How easy to customize
Iron	0.049 - 0.098	169	7.1 g/cc	Difficult
Aluminum	0.307 – 1.221	80	2.68 g/cc	Difficult
ABS	29.30	25-50	1.01-1.21 Mg/ m^3	Easy
PLA	29.30	36-55	1.25 Mg/ m^3	Easy

Table 3. Advantages & Disadvantages Comparison

Materials	Advantages	Disadvantages
Iron	• Strong • Durable	• Heavy Weight • Expensive • Hard to recycle • Not environment friendly • Rusts
Aluminum	• Light Weight • Durable • Won't Rust • Doesn't deteriorate	• Not very strong • Expensive • Hard to recycle • Not environment friendly • Can easily be water stained
3D Printer Materials	• Light Weight • Cheap • Easy to recycle • Environment Friendly • Flexible • Durable • Won't Rust • Doesn't deteriorate	• Not very strong

3.5 Electromechanical System

1) Battery:

For a battery, there are different kinds of batteries and each of them has its own features. First of them is Lithium ion Battery which has very high capacity, normal size, long life, high cost and it is rechargeable. Second one is NiMH which has a small size, low capacity, short life, low cost and non-rechargeable.

Third one is NiCd has big size, low capacity, short life, low cost, and rechargeable. Final one is SLA-6v20 has very big size, very high capacity, long life, high cost, and rechargeable. Energizer NH15BP and Duracell have short life and low capacity so they will not benefit this project well. Lithium ion and SLA-6v20 both are good, but we will choose Lithium ion because its size is smaller. Batteries comparison shown in Table 4.

2) Actuators:

Actuators are mechanical devices that move or turn energy into motion. There are different types of actuators, which are electric

actuators, hydraulic and pneumatic actuators. Electric actuators transfer electrical energy into mechanical energy. There are two types of electric actuators which are direct current motor (DC motor) and servomotors. Brushless DC motor is a type of DC motor that has different properties. It is small in size and weight not to mention its high speed and torque. On the other hand, a servomotor is a type of actuator that its functionality is to control motion. It operates within the limit of the specified angle (angular precision) by receiving a control signal to take further action. Moreover, it uses closed-loop feedback to control speed, torque or position. Pneumatic actuator is a hollow cylinder inside it a piston. The piston moves by applying a pressure from a pneumatic pump (compresses air) to create a force. Hydraulic actuators work similarly to Pneumatic actuators; except for the fact they are driven by liquid (fluid) pressure instead of air pressure. Pneumatic actuators are more expensive than hydraulic actuators. Moreover, hydraulic actuators are stronger or have greater force to move heavy loads.

Table 4. Batteries Types

Specifications	Li-Ion			NiMH	NiCd
	Cobalt	Phosphate	Manganese		
Energy (Wh/Kg)	150-250	100-150	90-120	60-120	45-80
Life Cycle (80% DoD)	500-1000	500-1000	1000-2000	300-500[3]	1000[3]
Charge Time (H)	2-4	1-2	1-2	2-4	1-2
Cell Voltage(v)	14.8	3.7	3.2-3.3	1.2	1.2
Safety Requirements	Protection circuit mandatory			Fuse protection, thermally stable	
Cost	High			Moderate	

3.6 Programming

3) Processor

Nowadays, the advancement in the processors and controller's technology has made it affordable and possible to achieve most of project requirements. In today's market, the top leading and the most common three processors/controllers are Raspberry pi, Arduino and Intel Edison. Raspberry pi considers as a tiny computer that has its own operating system. It is Linux based operating system and it can multitask. The board includes a processor, Ram, USB ports and everything a normal computer has. There are four models or generations of raspberry pi, which are Raspberry Pi 1, Raspberry Pi 2, Raspberry Pi 3, and Raspberry zero. Each one of them has different properties and using them depend on the needs. In general, a raspberry pi allows the user to not be restricted or limited to a certain programing language. Also, it is very fast in processing and easy to connect to Internet

On the other hand, Arduino is a microcontroller that is consists of software and hardware. It is a board that contains a chip to be programmed and use it to perform several functions. It is flexible that can interact with different devices like LEDs, buttons, speakers, motors, and cameras. For instance, it can read data from sensors, get to blink an LED, etc. There are several types of Arduino but the most popular one is the Arduino UNO. Most Arduinos have the same components, which are a power source, a processor, digital and analog pins, a reset button and a USB port. Arduino is easier to work with and it is better in controlling than Raspberry pi. It is capable of controlling complex external hardware. Also, it is cheaper than Raspberry pi. There is also another type of microprocessor that considers as a competitor to raspberry pi. It is offered by Intel Company, which is called Intel Edison. It functions as a computer. The board has a processor, RAM, USB port, Bluetooth and WiFi. It is a little bit similar to Raspberry pi but each one of them has its own features. For example, Intel Edison is more expensive than raspberry pi. Also, it has less numbers of USB ports and the processor speed is lower than the one in raspberry pi.

On a separate note, microcontrollers and microprocessors show different weaknesses and strengths depending on the application. There is a difference between microcontroller and microprocessor and each has advantages and disadvantages. In general, microcontrollers are more suitable for controlling devices (such as servos, etc.) whereas the microprocessors are better in processing

4) Interface

There are so many user interfaces to help the user interact with the computer to detect and analyze specific information, or to get some feedback from it. EEG is one of the user interfaces that measured by using electrodes that is placed on the scalp. It can be recorded by the electric fields that are generated by the nerve cells in the brain. EEG has so many advantages such as the characteristic of the electrical recording system because it has high precision and time measurements. Another advantage is that EEG is a very inexpensive device, and can easily be operated with. Poor recording spatial resolution is one of the disadvantages of the EEG. EMOTIV provides two different headsets, which are EPOC+ and Insight. Each one has different properties or features. The main differences that will serve our needs are related to the number of sensors and signal resolution. The EPOC features 16 sensors plus 2 reference sensors. Also, it provides high resolution. On the other hand, the insight has 5 sensors plus 2 reference sensors and it is less accurate than EPOC when detecting the signals (low resolution).

The second user interface is virtual reality. It is an artificial environment that is made by software and presented by the user in such a way that makes it real. Virtual reality also has so many advantages and disadvantages. One of the most important advantages is that the disabled people that are not able to experience reality can explore the virtual world, and experience the full life there. It also allows the user to experience impossible things in real life. One of the biggest drawback of the virtual reality is that people might get addicted to that virtual world which will lead them to forget their responsibilities in the real life.

The third user interface is the electromyography (EMG). It is way to access the health muscles and nerves to control and interact with it. EMG signals can be detected or transferred through a small device called electrodes. It has some advantages and disadvantages. EMG is more accurate, safe, and easy to get the wanted results. Also, EMG signals are not as complicated as the EEG when it comes to finding locations for measuring. However,

noises may occur due to the distance between the user and the signals of muscles.

3.7 Communication

The first category to be chosen is to have wireless communication to give the user the freedom to move while using the device. The Wifi, Bluetooth, and XBEE (pronounced ZigBee) are three different types of wireless communications which varies in the Safety, range of accessibility, reliability, power consumption, and bandwidth.

Starting with WiFi, its security is lower than the other two types with 20-150 meters range of communication. Although the reliability is low it still consumes high power with its wide bandwidth. Second choice is the Bluetooth which has a bit better safety communication than the Wifi. But the main problem is the short range which is 8-30 meters. Still it has low reliability but low power consumption with a narrow bandwidth. The third option is the ZigBee is offering a good safety communication with a good range of communication coverage 20-150 meters. It does have a high reliability compared to the other two options as well as low power consumption with a narrow bandwidth

3.8 Decision Making and Selections

Based on the project needs and requirements, a 3D printed design was used. The pressure sensor chosen to be LPS25MB because it is cheap and light in weight. Also MS5637-02BA03 was chosen to be the selected pressure sensor for its low cost and high accuracy. For the health sensor Pulse and Oxygen in Blood Sensor (SPO2) was chosen due to its accuracy and fast response. And the servo motor ASMC-03B was selected due to its high torque even it's a bit more expensive than other servo motors. Cobalt Li-ion battery got the highest score because of its suitable cell voltage, long life time, good energy, suitable charge time, low cost and it is safe. For the interface, EEG and EPOC allows obtaining more or variety of locations than insight. Based on the comparison, EPOC is more suitable than insight in terms of accuracy and sensors location. Finally, Arduino was chosen because it has the least price above them all. In case something happened during the coding process and the circuit got damaged we can easily buy another one because of its low price. Also, it can help us easily in our project since it has so many libraries and control system to control the exoskeleton.

4. Results and Analysis

The servo motors are responsible for the fingers movement according to the selected exercise. Based on the user selection the servo motors will perform the selected task. In other words, if the user chooses one the Arduino will give the corresponding signals for the motors to perform exercise one which is opening the hand five fingers then close them twice with a delay five seconds after each open or close process. The second exercise that might be selected by the user also is to open the fingers and closing them five times with the same delay time in between. The third and last exercise is opening and closing the fingers nine times. Beside this the medical sensors such as the temperature sensor is helping the assisting the user continuously check his body temperature. Another medical sensor mounted on the device is the pulse oximetry sensor to monitor the arterial oxygen saturation in the hemoglobin. Also a regular sixteen characters two lines LCD is used to show the measured and monitored data for the user with a Wi-Fi connected between the wearable device and the mobile phone application to send the collected data. Innovator X Post-Op Elbow was used as an angular controller. It provides safety to the user due to the adjustments of the angle position that we can make to prevent hand broken.

The biggest challenge of the project was to build a pattern recognition system for the brain signals to be able to detect the required movement from the brain itself with any other external method. The problem is that making a wrong pattern recognition may hurt the patient if a wrong move was done if the brain signal was not read correctly. A raw brain signal data coming from the EEG was read and analyzed by MATLAB software. MATLAB included a digital band pass and band stop filters to eliminate the noise from the brain signal.

5. Conclusion

Patients suffering from some kind of hand disabilities are facing a problem in their day to day activities that's why in this paper we've suggested a light weight low-cost wearable device compared to the devices used nowadays. The device designed to assist the user through two different modes of operation. The two modes are active and passive and both are helping the user to perform some hand exercises as well as monitoring his body temperature, pulses, ECG, and oxygen percentage.

Exoskeleton is a wearable device that improves the user's performance. There are two kinds of exoskeleton, which are assistive and rehabilitation. Assistive exoskeleton gives a user an extra strength to do different tasks. Rehabilitation exoskeleton reduces the recovery time and has long-term effect. This project will help many people that suffer from diseases like stroke, broken bones, spinal cord injury, hemiplegia and traumatic brain injury. Hands are so important part of the human body, so we will design a hand rehabilitation exoskeleton for stroke patients. This hand works in two modes: passive and active. It will be free size. This exoskeleton expected to increase the strength of the hand, the chance of healing, decrease the pain, the cost, and the recovery time.

Our design contains mainly four parts, first the EEG part, which will read the signals from the brain and send it to the PRU part through the Wi-Fi. PRU stands for pattern recognition unit that its function is to analyze signals. Signals were analyzed using Matlab. The third part, which is electromechanical part, will take the signals from PRU through ZigBee and control the arm servomotor using Arduino. Electromechanical part also contains five small servomotors to control the fingers to do specific exercises. Health care part contains three sensors, which are temperature sensor, pulse sensor, and ECG sensor and the results will be shown in the LCD. The user also can enter his information and choose any exercise using keypad and LCD. The last part is mobile application, which will take the results from the health care part through the Wi-Fi.

As a suggested future work and improvements the device can be improved by selecting the control mode through the brain signals with the keypad used as an extra input method. Also the keypad with the screen can be replaced by one big touch screen.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank the American University of the Middle East that gave us the opportunity to carry out our research. We would also like to thank the Robotics Research center for funding the project providing us such great facilities and assistance throughout our research.

References

[1] F Sayegh, F Fadhli, F Karam, M BoAbbas, F Mahmeed, JA Korbane, S AlKork, T Beyrouthy "A wearable rehabilitation device for paralysis," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp.1-4.

[2] A. P. Ong and N. T. Bugtai, "Recent Developments of Robotic Exoskeletons for Hand Rehabilitation," Presented at The DLSU Research Congress 2016, pp. 1-2.

[3] C. A. Lim 2013, "Ironman Project-Design Of Electro-Mechanical Muscle For Elbow Exoskeleton Robot," 2013.

[4] Rahman, M. A., & Al-Jumaily, A. Design and development of a hand exoskeleton for rehabilitation following stroke. Procedia Engineering, 41, 1028-1034, 2012.

[5] Kawasaki, H., Kimura, H., Ito, S., Nishimoto, Y., & Hayashi, H. Hand rehabilitation support system based on self-motion control, with a clinical case report. In *2006 World Automation Congress* (pp. 1-6). IEEE. (2006, July).

[6] Heo, P., Gu, G. M., Lee, S. J., Rhee, K., & Kim, J. Current hand exoskeleton technologies for rehabilitation and assistive engineering. *International Journal of Precision Engineering and Manufacturing*, *13*(5), 807-824. (2012).

[7] Sandoval-Gonzalez, O., Jacinto-Villegas, J., Herrera-Aguilar, I., Portillo-Rodiguez, O., Tripicchio, P., Hernandez-Ramos, M., ... & Avizzano, C. Design and Development of a Hand Exoskeleton Robot for Active and Passive Rehabilitation (2016).

[8] Ho, N. S. K., Tong, K. Y., Hu, X. L., Fung, K. L., Wei, X. J., Rong, W., & Susanto, E. A. An EMG-driven exoskeleton hand robotic training device on chronic stroke subjects: task training system for stroke rehabilitation. In *2011 IEEE international conference on rehabilitation robotics*(pp. 1-5). IEEE (2011, June).

[9] Babaiasl, M., Mahdioun, S. H., Jaryani, P., & Yazdani, M. A review of technological and clinical aspects of robot-aided rehabilitation of upper-extremity after stroke. *Disability and Rehabilitation: Assistive Technology*, *11*(4), 263-280 (2016).

[10] Arata, J., Ohmoto, K., Gassert, R., Lambercy, O., Fujimoto, H., & Wada, I. A new hand exoskeleton device for rehabilitation using a three-layered sliding spring mechanism. In Robotics and Automation (ICRA), 2013 IEEE International Conference on (pp. 3902-3907). IEEE (2013,May).

[11] Sarakoglou, I., Tsagarakis, N. G., & Caldwell, D. G. (2004, September). Occupational and physical therapy using a hand exoskeleton based exerciser. In Intelligent Robots and Systems. (IROS 2004). Proceedings. IEEE/RSJ International Conference on (Vol. 3, pp. 2973-2978). IEEE) (2004).

[12] Capitani, S., Cremoni, A., Lindenroth, L., Secciani, N., Shaftr, A., Stilli, A., &Venture, M. Development of low-cost portable hand exoskeleton for assistive and rehabilitation purposes. For The ENTERFACE International Workshop (pp. 1-3) (2016).

[13] W. D. Hairston, K. W. Whitaker, A. J. Ries, J. M. Vettel, J. C. Bradford, S. E. Kerick, and K. McDowell, "Usability of four commercially-oriented EEG systems," in *Journal of neural engineering*, *11*(4), 046018, 2014.

[14] abilah Hamzah, Haryanti Norhazman, Norliza Zaini and Maizura Sani. Classification of Eeg Signals Based on Different Motor Movement Using Multi-layer Perceptron Artificial Neural Network. *Journal of Biological Sciences, 16: 265-271* (2016).

[15] F. Lotte, L. Bougrain, and M. Clerc, "Electroencephalography (EEG)-based Brain Computer Interfaces," Wiley Encyclopedia of Electrical and Electronics Engineering, Wiley, pp. 44, 2015.

[16] F. Lotte, "A Tutorial on EEG Signal Processing Techniques for Mental State Recognition in Brain-Computer Interfaces," Eduardo Reck Miranda; Julien Castet. Guide to Brain-Computer Music Interfacing, Springer, 2014.

[17] M. M. Moazzami, "*EEG signal Processing in Brain-Computer Interface"* (Doctoral dissertation, Michigan State University), 2012.

[18] S AlAwadhi, N AlHabib, D Murad, F AlDeei, M AlHouti, T Beyrouthy, S Al-Kork "Virtual reality application for interactive and informative learning," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4.

[19] S.A. Awadhi, N.A. Habib, D. Al-Murad, F.A. deei, M.A. Houti, T. Beyrouthy, S. Al-Kork "Interactive Virtual Reality Educational Application", Advances in Science, Technology and Engineering Systems Journal, vol. 3, no. 4, pp. 72-82 (2018).

[20] Reaz, M. B. I., Hussain, M. S., & Mohd-Yasin, F. Techniques of EMG signal analysis: detection, processing, classification and applications. Biological procedures online, 8(1), 11-35 (2006).

[21] S. Said, S. AlKork, T. Beyrouthy and M. F. Abdrabbo, "Wearable bio-sensors bracelet for driver as health emergency detection," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSmart), Paris, 2017, pp. 1-4.

[22] Hasan S., Al-Kandari K., Al-Awadhi E., Jaafar A., Al-Farhan B., Hassan M., Said S. and AlKork S. (2018). Wearable Mind Thoughts Controlled Open Source 3D Printed Arm with Embedded Sensor Feedback System.In Proceedings of the 2nd International Conference on Computer-Human Interaction Research and Applications - Volume 1: CHIRA, ISBN 978-989-758-328-5, pages 141-149.

[23] S. K. A. Kork, I. Gowthami, X. Savatier, T. Beyrouthy, J. A. Korbane and S. Roshdi, "Biometric database for human gait recognition using wearable sensors and a smartphone," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4.

[24] D. Planelles, E. Hortal, Á. Costa, A. Ubeda, E. Iaez, and J. M. Azorin, "Evaluating classifiers to detect arm movement intention from eeg signals," *Sensors*, *14*(10), pp. 18172-18186, 2014.

[25] Taha Beyrouthy, Samer Al. Kork, J. Korbane, M. Abouelela, "EEG Mind Controlled Smart Prosthetic Arm – A Comprehensive Study", Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 3, pp. 891-899 (2017).

[26] T. Beyrouthy, S. K. Al Kork, J. A. Korbane and A. Abdulmonem, "EEG Mind controlled Smart Prosthetic Arm," 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 404-409.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 17-22 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

High-Temperature Optical Characterization of Wide Band Gap Light Emitting Diodes and Photodiodes for Future Power Module Application

Syam Madhusoodhanan[1], Abbas Sabbar[1], Sattar Al-Kabi[1], Stanley Atcitty[2], Robert Kaplar[2], Binzhong Dong[3], Jiangbo Wang[3], Shui-Qing Yu[1], Zhong Chen[*,1]

[1]*Department of Electrical Engineering, University of Arkansas, Fayetteville, Arkansas 72701, USA*

[2]*Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*

[3]*HC SemiTek (Suzhou), 28 Chenfeng Road, Zhangjiagang, Jiangsu, P. R. 215600, China*

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 22 February, 2019
Online : 06 March, 2019

Keywords:
High-Density Power Module,
High-Temperature Optocouplers,
Light Emitting Diodes,
Optical Characterization,
Photodiodes,
Wide Bandgap

ABSTRACT

A systematic study of wide bandgap (WBG) based light emitting diodes (LEDs) and photodiodes (PDs) were conducted for the assessment of modular integration of optoelectronic devices into power modules. The temperature dependence of the photoluminescence (PL) efficiency of Indium gallium nitride/Gallium nitride (InGaN/GaN) multiple quantum wells (MQWs) material was studied from 10 to 800 K. The photoluminescence efficiency is calculated using the power law relation connecting the integrated photoluminescence signal and the excitation pump power. A peak PL efficiency of 43.97 % was recorded at 800 K. Electroluminescence (EL) study was conducted on Gallium nitride (GaN) based LEDs in the temperature range of 300 – 800 K. An intensity drop of three orders of magnitude is recorded at 800 K as compared to room temperature. The full width half maximum (FWHM) of the spectra was also calculated from the EL measurements. The S-shaped shift of FWHM at high temperature indicates a strong localization effect. High-temperature spectral response analysis of 4H-Silicon carbide (4H-SiC) pn-junction photodiode is performed at zero voltage bias condition. Enhancement of spectral response is observed at higher wavelengths due to increased phonon population at higher temperatures.

1. Introduction

This paper is an extension of work originally presented in Conference on Lasers and Electro-Optics (CLEO): Applications and Technology 2018 titled 'Investigation of High-Temperature Photoluminescence Efficiency from InGaN/GaN MQWs' [1]. Power electronic system plays a significant role in renewable energy, energy storage, and electric or hybrid electric vehicles. As an integral part of the power electronic system, power modules provide a highly efficient form of power conversion. In the last two decades, the field of power conversion is going through a transformation where the application often demands extreme operating environments. Advancement in the field of hybrid electric vehicles, aerospace and deep oil-gas exploration necessitate the development of power modules based on WBG devices to operate in extreme environments where the ambient temperature exceeds 200oC. Compared to silicon (Si) devices, wide bandgap materials like SiC and GaN can withstand higher current densities, exhibit lower switching losses and possess higher breakdown voltages [2-3].

Power modules based on WBG materials enhance reliability and considerably reduce cooling requirements that lead to a significant reduction in total system cost and weight. Although these innovative properties lead power modules to higher power density [4], some concerns still need to be addressed to take full advantage of wide bandgap based modules. For example, the use of bulky transformers used as a galvanic isolation system to float the high voltage gate driver limit further size reduction of the high-temperature power modules [5]. Bulky transformers can be replaced by integrating high-temperature optocouplers to scale down power modules further and achieve disrupting performance

[*]Zhong Chen, University of Arkansas, chenz@uark.edu

in terms of thermal management, power efficiency, power density, operating environments, and reliability. However, regular semiconductor optoelectronic materials and devices have major difficulty functioning in high-temperature environments. The motivation for this study is to develop optoelectronic devices, specifically optocouplers that can be integrated into high-density power modules. A detailed study on optoelectronic devices at high temperature enables us to explore the possibility of scaling high-density power modules by integrating high-temperature optoelectronic devices into the power module.

Modular integration of optoelectronic devices into high-temperature power modules is restricted due to the significant optical efficiency drop at elevated temperatures. The quantum efficiency and long-term reliability of optoelectronic devices decrease at elevated temperatures. Thermal studies show that high junction temperature will significantly lower the lifetime of the LED [6-7]. Performance, as well as the useful life period of the LEDs, will significantly go down under high temperatures. It was also found that high leakage current [8-13], dark spot generation [10,11], and degradation of the metal contacts [9,12,13] at high temperatures can often lead to the degradation of light intensity. These factors limit the application of LEDs in high-temperature environments.

Photodiodes (PDs) based on SiC has lower leakage current and better stability at high-temperature operation. Brown et al. demonstrated a 6H-SiC pn-junction based PD operation at temperatures up to 350oC [13]. Several groups have reported 6H-SiC pn-junction photodiodes in the past [15-17]. Schottky-barrier PDs are also tested for high-temperature environments in recent years. Different groups have studied and characterized 4H-SiC Schottky-barrier and pn-junction PDs at high temperatures [18,19]. Although high quantum efficiency was reported up to 350oC, no groups have ever reported a detailed study on the temperature dependence of the spectral response of SiC pn-junction photodiodes for a wide range of temperatures. In this paper, we report on the performance degradation of InGaN based LED material and LED devices at high temperatures. Also, a study on the temperature dependence of the response in SiC-based pn-junction PDs up to 527oC (800 K) is presented.

2. Experiment

High-temperature photoluminescence studies were conducted on the InGaN/GaN MQW structure as described in the sketch of Fig. 1. The InGaN/GaN MQW structures were grown on sapphire substrates. The quantum well structure comprised of 10 periods of 3 nm undoped InGaN wells and 12.5 nm Si-doped GaN barriers grown on a 0.4 µm thick n-GaN layer. Vertically structured GaN-based bare die blue LEDs, fabricated at Cree, were subjected to high-temperature electroluminescence studies. The LED structures were grown on SiC substrate with both contact layers were formed by highly reflective electrodes. Optical responsivity measurements were conducted on 4H-SiC pn-junction PDs purchased from ifw Optronics GmbH. The material and device selection for this study are based on the demonstrated high quality, performance, and reliability.

Temperature-dependent measurements on materials and devices were performed using a Janis ST-100 cryostat. The cryostat can reach temperatures up to 800 K. Temperature-dependent PL measurements were carried out using a customary off-axis configuration. The excitation source was a continuous wave laser operating at 395 nm laser beam diameter of 64µm. The power instance of the sample was changed from 1.0 – 110.0 mW. A Horiba 550 spectrometer combined with a photomultiplier tube is used to collect PL and EL spectra. A Lakeshore 335 temperature controller is used to vary the temperature. Vacuum environment is maintained inside the cryostat throughout the measurements using a Janis TS-75-W turbopump. The injection current during the EL measurements was limited to 1.0 mA to avoid the device burn out due to the high current density. A Keithley 2450 source measurement unit is used for precise injection of current to the device. Temperature-dependent optical responsivity measurements on 4H-SiC PDs were carried out using Spex 270M rapid scanning imaging monochromator coupled with a deuterium lamp. The wavelength was changed from 250 to 400 nm.

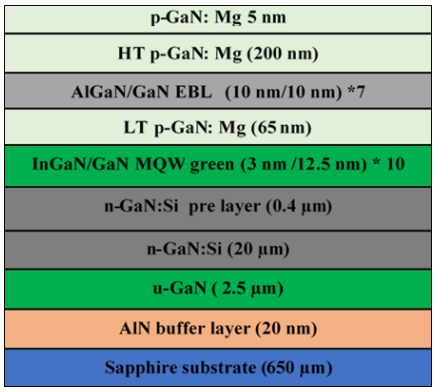


Figure 1. A sketch represents the InGaN/GaN MQWs structure.

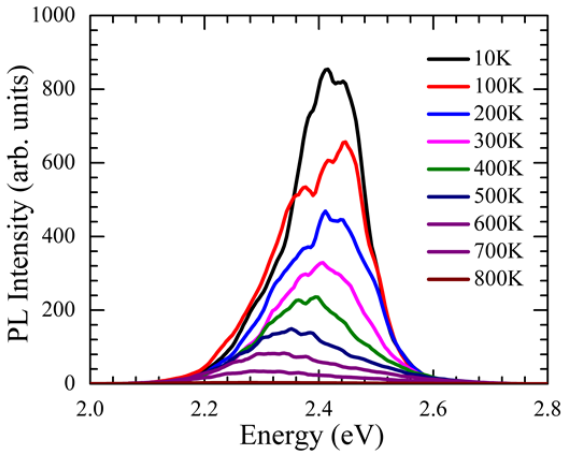


Figure 2. Temperature-dependent photoluminescence spectra of InGaN/GaN MQW were taken in the temperature range of 10 – 800 K.

3. Results and Discussion

The photoluminescence efficiency of the InGaN/GaN MQW structure was measured using the power law relation connecting the integrated PL intensity and the excitation pumping power of the spectrum [20]. Figure 2 shows PL spectra for the InGaN/GaN MQWs for a temperature range from 10 to 800 K. A gradual reduction in the intensity is observed on the high energy side of the PL spectra at elevated temperatures. It is observed from the plot

that the PL band of InGaN/GaN MQWs become wider at elevated temperatures. A strong Indium (In) segregation can be attributed to the temperature independent behavior of the low energy side [21]. The strong In segregation leads to the formation of In-rich quantum dot-like regions [22]. The spectral peak positions were redshifted from 2.41 eV at 10 K to 2.32 eV at 800 K. Varshni empirical equation explain the redshift due to the bandgap shrinkage [23].

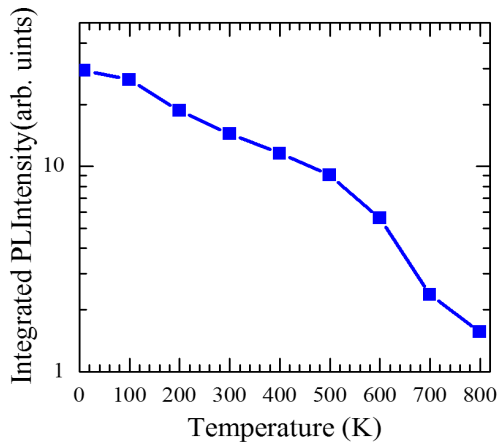


Figure 3. Integrated photoluminescence intensity of InGaN/GaN MQW is plotted as a function of temperature.

Figure 3 shows the integrated PL intensity for the InGaN-related PL emission for different temperatures. A reduction of one magnitude in the integrated PL intensity at 800 K compared to 77 K suggests a high PL efficiency even at elevated temperatures. Domination of non-radiative recombination at higher temperature causes a reduction in PL intensity. Carrier escape at higher temperature due to high thermal energy and the high rate of non-radiative recombination in QWs lead to the intensity drop at high temperature.

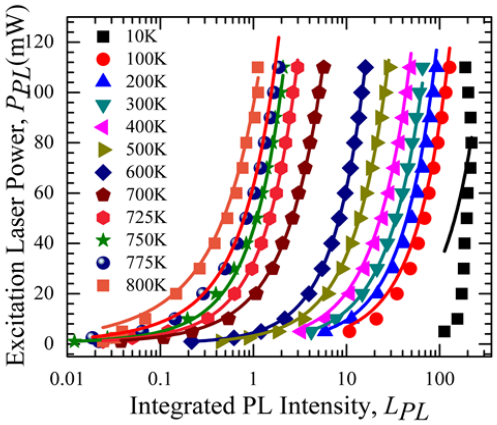


Figure 4. Excitation laser power is plotted against integrated PL intensity from InGaN/GaN MQW at different temperatures. The solid curves are fits of power law model to the data.

The extraction of PL efficiency by the power-law equation is explained in Ref. [19]. Excitation power dependent PL measurements were taken with pump power ranging from 1.0 mW to 110.0 mW. The results are plotted in terms of excitation pump power as a function of integrated PL intensity in Fig. 4. The dots are the experiment results while the solid curves are the fits of Eq. (1) from Ref. [19] to the data.

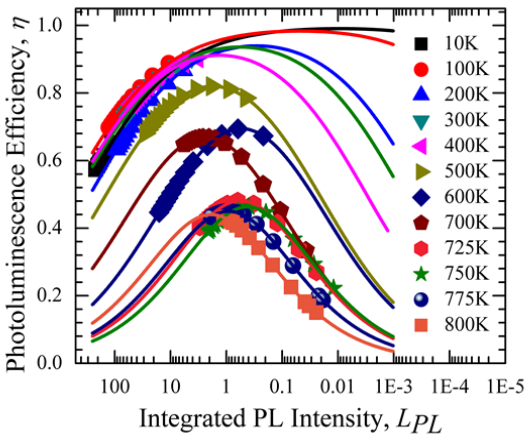


Figure 5. Photoluminescence efficiency is plotted as a function of integrated PL intensity at different temperatures. The solid lines are the simulated data for injection levels above and below the experiment limitations.

Figure (5) shows the temperature-dependent PL efficiency versus integrated PL intensity. The laser injection level where the peak PL efficiency is observed increases as the temperature increases. At 800 K, the peak PL efficiency is observed at an injection level above 110.0 mW. This behavior indicates that the non-radiative recombination rates are dominant even at a low injection rate under room temperature. At elevated temperatures, non-radiative recombinations along with thermally activated carrier escape from the QWs cause the PL efficiency to drop. These results indicate that the Auger and Shockley-Read-Hall (SRH) recombination rates are dependent on temperature and injection level.

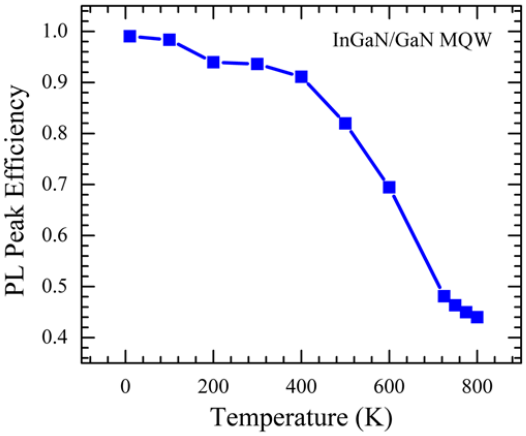


Figure 6. PL peak efficiency of InGaN/GaN MQW is plotted as a function of temperature.

The peak PL efficiency was plotted against temperature as shown in Fig. 6. The peak PL efficiency is almost unity at 10 K. At elevated temperatures, the peak PL efficiency decreases. As mentioned earlier, the drop in the Pl efficiency is caused by the domination of nonradiative recombinations in the active region of the quantum wells. At high temperatures, both SRH and Auger recombinations suppress the radiative recombination resulting in a PL efficiency drop. Another reason for the drop in PL efficiency is due to the carrier escape from the quantum wells as a result of high

thermal energy. A PL efficiency of 44% is observed for the sample at 800 K. The temperature independent behavior of PL efficiency up to 400 K indicates high thermal stability.

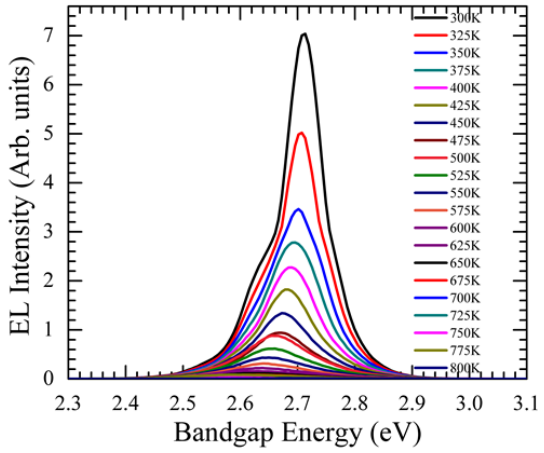


Figure 7. Electroluminescence spectra of InGaN/GaN single quantum well LED measured at different temperatures ranging from 300 – 800 K.

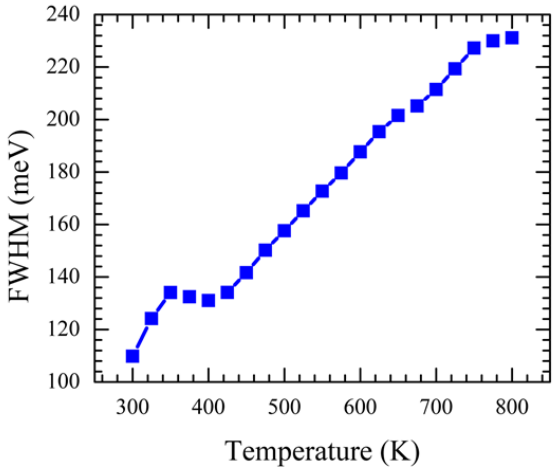


Figure 8. The full width half maximum of the electroluminescence spectra is calculated and plotted against temperature for InGaN/GaN single quantum well LED from 300 – 800 K.

Temperature-dependent electroluminescence spectra of the InGaN/GaN single quantum well (SQW) LED at an injection current of 1 mA is shown in Fig. 7. The spectral intensity of the LED is decreased as the temperature is increased from 300 to 800 K. The EL peak energy is measured to be 2.71 eV at 300 K. The Indium composition is approximated to be as 16 % from the EL peak energy [24]. The EL peak energy changes from 2.72 to 2.60 eV as the temperature increased from 300 to 800 K. The EL spectra are fitted using a Gaussian model to calculate the full-width half maximum (FWHM). It is found that the FWHM of the spectra increased from 109.11 meV at 300 K to 230.11 meV at 800 K. Temperature dependence of the FWHM of the EL spectra for InGaN/GaN LED is shown in Fig. 8. The FWHM shows an S-shaped shift when the temperature increases. The S-shaped shift again indicates a strong localization effect [25].

The FWHM of the spectra increased from 109.11 meV at 300 K to 230.11 meV at 800 K. The FWHM increases from 300 - 350 K due to carrier relaxation in the localized states. The further increase in the temperature results in a drop in the FWHM until the temperature reaches 400 K. The drop in the FWHM when the temperature further increases to 400 K suggests that nonradiative recombination centers capture a part of the carriers before reaching the lower energy tail states. The FWHM increases when the temperature increases above 400 K. This monotonic increase in FWHM is caused by the coupling of the excitons to acoustic phonons and longitudinal optical (LO) phonons [26].

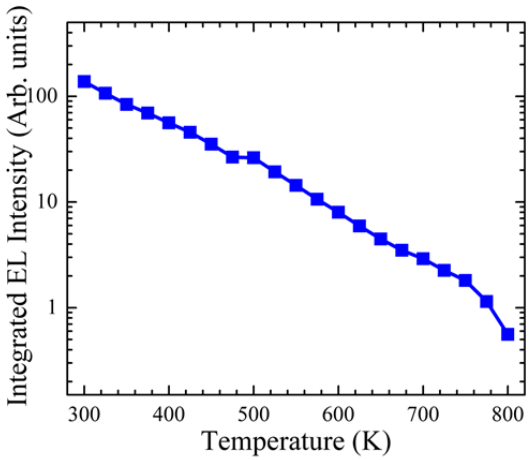


Figure 9. The integrated intensity of the electroluminescence spectra at each temperature is calculated and plotted as a function of temperature for InGaN/GaN single quantum well LED.

Figure 9 shows the change in the integrated intensity of the InGaN/GaN SQW LED over the temperature range of 300 – 800 K. The integrated intensity decreases when the temperature increases. The integrated intensity remains the same in the temperature range of 475 – 500 K. An intensity drop of three orders of magnitude is recorded at 800 K as compared to room temperature. The three orders of magnitude drop in the intensity suggest the presence of shallow QWs in the InGaN/GaN structure. The reduced intensity at a high temperature is due to the domination of the nonradiative recombination. At elevated temperature, the injected carriers gain high thermal energy. Carrier escape due to high thermal energy leads to the intensity drop at high temperature.

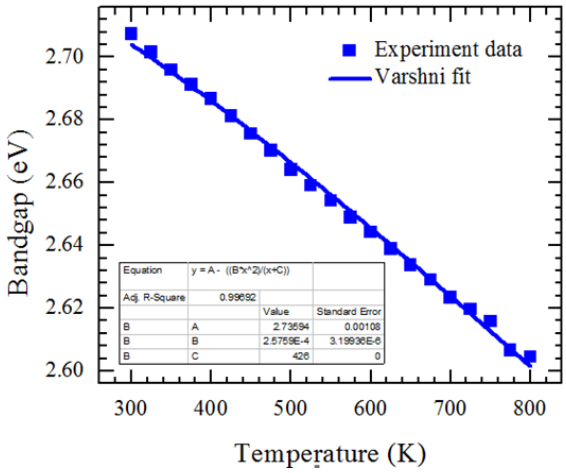


Figure 10. Energy bandgap of InGaN/GaN single quantum well LED is measured and plotted against temperature.

Figure 10 shows the change in the EL peak energy of the InGaN/GaN SQW LED concerning temperature. The EL peak energy decreases when the temperature increases. The rate of change in EL peak energy remains almost constant during the temperature range of 300 – 750 K. The EL peak energy changes from 2.71 to 2.60 eV over the temperature range of 300 - 800 K. The temperature dependence of the EL peak energy is described by the Varshni relation [23].

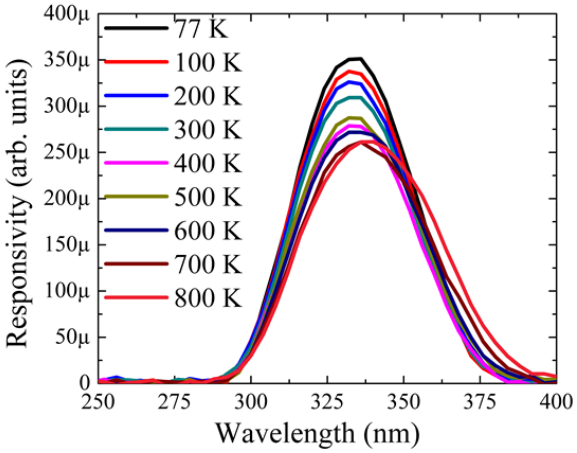


Figure 11. The spectral response characteristics of 4H-SiC pn-junction photodiode at different temperatures under 0 V bias.

Figure 11 shows the spectral response characteristics of the 4H-SiC detector measured as a function of temperature from 77 K to 800 K at zero biased condition. The photodetectors spectral response is wavelength dependent on the incident light. This is caused by a different energy absorption rate at each wavelength. The spectral response curves shift towards longer wavelength, and the responsivity at a wavelength above 350 nm is improved at higher temperatures. The redshift is caused by thermally induced bandgap narrowing effect. The observed spectral enhancement at higher wavelengths can be attributed to the increased phonon population at higher temperatures. Phonons assisted photon absorption is necessary at longer wavelengths as the corresponding photon energy is smaller than the direct energy spacing of 4H-SiC [19]. Interestingly, the responsivity at a lower wavelength (<320nm) exhibits less temperature dependent. The responsivity spectra of the photodiode at lower wavelength merges indicating that at this region temperature has less influence on the responsivity. More studies are necessary to understand the degradation of detectors under high temperature thoroughly.

4. Conclusion

Performance degradation of GaN-based LED material, and devices are studied for a wide range of temperature. Spectral responsivity of a 4H-SiC pn-junction photodiode is evaluated for the temperature range of 77 -800 K. The InGaN/GaN MQW structure shows temperature independent PL efficiency up to 400 K, signifying high thermal stability and displays a peak PL efficiency of 44 % at 800 K. The EL intensity of the LED device was decreased by three orders of magnitude at high temperatures as compared to room temperature. The reduced EL intensity is caused by the carrier escape from the quantum well due to high thermal energy. A strong localization effect is speculated due to the S-shaped FWHM curve. Initial studies on the 4H-SiC pn-junction photodiode show a temperature independent spectral response at lower wavelengths. These results serve as a strong foundation for future studies in the field of high-temperature optoelectronics in the applications of future ultra-high density power modules.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors acknowledge financial support from the United States Department of Energy, DOE (Grant No. DE-SC0016485), program manager Dr. Imre Gyuk. Sandia National Laboratories is a multi-program laboratory managed and operated by National Technology Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE- NA0003525. The views expressed in this article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] Sabbar, A., Madhusoodhanan, S., Al-Kabi, S., Dong, B., Wang, J., Yu, S., & Chen, Z. (2018). Investigation of High-Temperature Photoluminescence Efficiency from InGaN/GaN MQWs. Conference on Lasers and Electro-Optics. doi:10.1364/cleo_at.2018.jth2a.79

[2] Homberger, J., Lostetter, A., Olejniczak, K., Mcnutt, T., Lal, S., & Mantooth, A. (n.d.). Silicon-carbide (SiC) semiconductor power electronics for extreme high-temperature environments. 2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No.04TH8720). doi:10.1109/aero.2004.1368048

[3] Nomura, T., Masuda, M., Ikeda, N., & Yoshida, S. (2008). Switching Characteristics of GaN HFETs in a Half Bridge Package for High Temperature Applications. IEEE Transactions on Power Electronics, 23(2), 692-697. doi:10.1109/tpel.2007.915671

[4] Katsis, D. C., & Zheng, Y. (2008). Development of an extreme temperature range silicon carbide power module for aerospace applications. 2008 IEEE Power Electronics Specialists Conference. doi:10.1109/pesc.2008.4591943

[5] Rowden, B., Mantooth, A., Ang, S., Lostetter, A., Hornberger, J., & Mcpherson, B. (2009). High Temperature SiC Power Module Packaging. Volume 5: Electronics and Photonics. doi:10.1115/imece2009-12883

[6] Narendran, N., & Gu, Y. (2005). Life of LED-Based White Light Sources. Journal of Display Technology, 1(1), 167-171. doi:10.1109/jdt.2005.852510

[7] Huang, F., Yin, L., Li, S., Xu, G., Yan, H., Chen, Y., . . . Zhang, J. (2011). Study on the reliability of high-power LEDs under temperature cycle. 2011 12th International Conference on Electronic Packaging Technology and High Density Packaging. doi:10.1109/icept.2011.6067024

[8] Cao, X., Sandvik, P., Leboeuf, S., & Arthur, S. (2003). Defect generation in InGaN/GaN light-emitting diodes under forward and reverse electrical stresses. Microelectronics Reliability, 43(12), 1987-1991. doi:10.1016/j.microrel.2003.06.001

[9] Osiński, M., Zeller, J., Chiu, P., Phillips, B. S., & Barton, D. L. (1996). AlGaN/InGaN/GaN blue light emitting diode degradation under pulsed current stress. Applied Physics Letters, 69(7), 898-900. doi:10.1063/1.116936

[10] Cao, X., Teetsov, J., Shahedipour-Sandvik, F., & Arthur, S. (2004). Microstructural origin of leakage current in GaN/InGaN light-emitting diodes. Journal of Crystal Growth, 264(1-3), 172-177. doi:10.1016/j.jcrysgro.2004.01.031

[11] Egawa, T., Jimbo, T., & Umeno, M. (1997). Characteristics of InGaN/AlGaN light-emitting diodes on sapphire substrates. Journal of Applied Physics, 82(11), 5816-5821. doi:10.1063/1.366450

[12] Bulashevich, K. A., Evstratov, I. Y., Mymrin, V. F., & Karpov, S. Y. (2007). Current spreading and thermal effects in blue LED dice. Physica Status Solidi (c), 4(1), 45-48. doi:10.1002/pssc.200673502

[13] Cao, X., Stokes, E., Sandvik, P., Taskar, N., Kretchmer, J., & Walker, D. (2002). Optimization of current spreading metal layer for GaN/InGaN-based light emitting diodes. Solid-State Electronics, 46(8), 1235-1239. doi:10.1016/s0038-1101(02)00023-0

[14] Brown, D., Downey, E., Ghezzo, M., Kretchmer, J., Saia, R., Liu, Y., . . . Schneider, W. (1993). Silicon carbide UV photodiodes. IEEE Transactions on Electron Devices, 40(2), 325-333. doi:10.1109/16.182509

[15] Campbell, R., & Chang, H. (1966). Detection of ultraviolet radiation using silicon carbide P-N junctions. 1966 International Electron Devices Meeting. doi:10.1109/iedm.1966.187651

[16] Glasow, P., Ziegler, G., Suttrop, W., Pensl, G., & Helbig, R. (1988). SiC-UV-Photodetectors. Optoelectronic Technologies for Remote Sensing from Space. doi:10.1117/12.943598

[17] Edmond, J., Kong, H., Suvorov, A., Waltz, D., & Carter, J. C. (1997). 6H-Silicon Carbide Light Emitting Diodes and UV Photodiodes. Physica Status Solidi (a), 162(1), 481-491. doi:10.1002/1521-396x(199707)162:13.0.co;2-o

[18] Watanabe, N., Kimoto, T., & Suda, J. (2012). 4H-SiC pn Photodiodes with Temperature-Independent Photoresponse up to 300oC. Applied Physics Express, 5(9), 094101. doi:10.1143/apex.5.094101

[19] Blank, T. V., Goldberg, Y. A., Kalinina, E. V., Konstantinov, O. V., Konstantinov, A. O., & Hallén, A. (2005). Temperature dependence of the photoelectric conversion quantum efficiency of 4H–SiC Schottky UV photodetectors. Semiconductor Science and Technology, 20(8), 710-715. doi:10.1088/0268-1242/20/8/010

[20] Johnson, S. R., Ding, D., Wang, J., Yu, S., & Zhang, Y. (2007). Excitation dependent photoluminescence measurements of the nonradiative lifetime and quantum efficiency in GaAs. Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures, 25(3), 1077. doi:10.1116/1.2720864

[21] Shchekin, O. B., Epler, J. E., Trottier, T. A., Margalith, T., Steigerwald, D. A., Holcomb, M. O. et al Krames, M. R. (2006). High performance thin-film flip-chip InGaN–GaN light-emitting diodes. Applied Physics Letters, 89(7), 071109. doi:10.1063/1.2337007

[22] Bogdanov, M. V., Bulashevich, K. A., Khokhlev, O. V., Evstratov, I. Y., Ramm, M. S., & Karpov, S. Y. (2010). Current crowding effect on light extraction efficiency of thin-film LEDs. Physica Status Solidi (c), 7(7-8), 2124-2126. doi:10.1002/pssc.200983415

[23] Y. P. Varshni, ``Temperature dependence of the energy gap in semiconductors," Physica, (Utrecht), no. 34, pp. 149 -154, 1967, http://dx.doi.org/10.1016/0031-8914(67)90062-6

[24] Wu, J., Walukiewicz, W., Yu, K. M., Ager, J. W., Haller, E. E., Lu, H., & Schaff, W. J. (2002). Small band gap bowing in In1−xGaxN alloys. Applied Physics Letters, 80(25), 4741-4743. doi:10.1063/1.1489481

[25] Lu, T., Ma, Z., Du, C., Fang, Y., Wu, H., Jiang, Y., . . . Chen, H. “Temperature-dependent photoluminescence in light-emitting diodes, Scientific Reports, 2014, http://dx.doi.org/10.1038/srep06131

[26] Franssen, G., Litwin-Staszewska, E., Piotrzkowski, R., Suski, T., & Perlin, P. (2003). Optical and electrical properties of homoepitaxially grown multiquantum well InGaN/GaN light-emitting diodes. Journal of Applied Physics, 94(9), 6122-6128. doi:10.1063/1.1616637

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 23-33 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Design of an Additively Manufacturable Multi-Material Light-Weight Gripper with integrated Bellows Actuators

Gabriel Dämmer[1,2,*], Sven Gablenz[1], Alexander Hildebrandt[1], Zoltan Major[2]

[1]*Festo AG & Co. KG, Advanced Mechatronic Systems, 73734, Germany*

[2]*Johannes Kepler University Linz, Institute of Polymer Product Engineering, 4040, Austria*

A R T I C L E I N F O

Article history:
Received: 22 December, 2018
Accepted: 09 February, 2019
Online: 06 March, 2019

Keywords:
Printed Robotics
Multi-Material Additive Manufacturing Structural Optimization
Pneumatic Actuators
Pneumatic Robots
Functional Integration
PolyJet™

A B S T R A C T

Combining state-of-the-art additive manufacturing technologies with structural optimization has the potential to produce geometrically complex multi-material components with integrated functionalities and desired structural behavior. In this article, the simulation-driven design process of a multi-material light-weight gripper with an integrated pneumatic bellows actuator is described. The design of the bellows structure is based on a previously published contribution to the RoboSoft2018 conference in Livorno, Italy. The conference paper contains the shape optimization and experimental investigations of the structural and fatigue behavior of linear type multi-material PolyJet bellows actuators. In this extended version, the main findings of the conference paper are translated into the design of a rotary type bellows actuator that is finally integrated into a multi-material light-weight gripper. In order to define the lay-out of the gripper's support structure, a density-based topology optimization is performed and the application on a PolyJet printed light-weight robot is demonstrated. The presented design approach and results are useful for researchers and engineers involved in the development of multi-material additive manufacturing, simulation-driven design and functionally integrated structures for pneumatic robotic systems.

1. Introduction

Recent advances in material science and manufacturing technologies allows for the production of geometrically complex [1], [2] multi-material structures that were not manufacturable until recently [3]. At the same time, an increasing demand for short lead times, large product variety and flexible production processes will influence the design of future robotic systems. Particularly interesting components for future light-weight robots may result from the combination of multi-material additive manufacturing (AM) and pneumatic actuation.

1.1. "Low Inertia and Compliant Elements" Approach

An approach for increased flexibility in future production processes is to combine human and robotic strengths in human-robot collaboration scenarios. The associated abolition of safety cages—that typically enclose robot workspaces today—leads to alternative safety concepts. In general, various measures can be applied to reduce the severity of injuries or damage caused by an unintended collision [4]. Thereby, the addition of compliant elements to the kinematic chain and the reduction of manipulator link inertia contributes to an inherently safe design [4], [5]. In the described work, novel materials and manufacturing technologies are combined with structural optimization to create compliant and light-weight components for light-weight robots.

1.2. Topology Optimization and Additive Manufacturing

The term topology optimization (TO) covers mathematical methods that allow for the definition of an optimum material distribution (also "material lay-out" [6]) within a given design space. Typical optimization set-ups are the minimization of structural compliance (or "global stiffness", [6]) with a constraint volume or the minimization of used material (volume) respecting a stress constraint. The underlying concept and fundamentals of todays TO have already been proposed in the 1990s (see [7], [8]).

* Gabriel Dämmer, Festo AG & Co.KG, gabriel.daemmer@festo.com

However, the field of TO has recently attracted a lot of attention. This actuality may (at least partly) be explained by an increasing degree of practical applicability of the optimization results through additive manufacturing of single material [9], [10], [11] as well as multi-material structures [3], [12]. Most AM processes allows for very complex geometries and therefore facilitate simulation-driven design processes. Thereby, light-weight design is especially relevant for robotic components close to the payload as their motion results in comparatively large inertial loads which in turn reduce the robot's performance and safety. Hence, considering AM technologies for the development of robot components makes sense in many ways [13]. Moreover, users of optimization approaches based on a given material layout, such as shape optimization (SO) and sizing optimization [14], can as well benefit from the absence of conventional manufacturing constraints as a result of the layer-wise material deposition inherent to many AM technologies.

1.3. Pneumatic Bellows Actuators

As a result of high gear ratios, typical electro-mechanical drive systems show little compliance when subjected to abrupt external loadings. Against the background of human-robot collaboration, mechanical designs have been developed that overcome this behavior i.e. can reduce the stiffness of electro-mechanical drive systems by the addition of compliant elements between the motor and torque output [15]. The use of nonlinear springs even allows for adjustment of the stiffness level [16], [17] so that the compliance of the system can be tuned according to a specific application. However, these measures come along with an inevitable increase in weight and mechanical complexity. The use of pneumatic actuators presents an alternative approach. Due to the relatively low compression modulus of air, pneumatic actuators show inherently compliant behavior and can easily be arranged to antagonistic pairs with adjustable compliance [17], [18], [19]. However, conventional pneumatic actuators (e.g. pneumatic cylinders) include dynamic seals and require smooth surfaces and close manufacturing tolerances [20]. Considering AM for the production of pneumatic actuators, alternative concepts such as bellows actuators are more promising. The structural behavior of bellows actuators can easily be modified by shape and material variations. Bellows actuators have been realized by various AM technologies such as selective laser sintering [21], PolyJet™ printing [20], [22] and Digital Mask Projection Stereolithography [23]. Numerous publications demonstrate the silicone molding process of bellow-like bending actuators utilizing additively manufactured molds (e.g. [24]). Also, detailed reviews of bellows actuators in the context of articulated robotic systems [25], 3D printing [26] and soft robotics [27] have been published. In continuum manipulators such as the "Bionic Handling Assistant" [21], bellows structures are not only used for actuation but also as a support structure.

1.4. The DIMAP Project

The main objective of the EU founded research project "Digital Materials for 3D Printing" (DIMAP) was the development of novel functional materials for PolyJet™ technology (http://www.dimap-project.eu). In PolyJet™ printing, three-dimensional objects are created from layers of small droplets of resin that are cured via ultra-violet light. Using this technology, multi-material structures with highly complex geometries can be manufactured. In order to narrow down the possible applications of the materials and technology in industry-like environments, a PolyJet-printable light-weight robot was developed.

1.5. Aim and Structure of this article

PolyJet™ printing is an interesting option for the manufacturing of highly geometrically complex and functionally integrated multi-material structures. These capabilities may be used to produce robotic components with desired structural behavior such as maximum stiffness (for light-weight structures) and inherent compliance (for actuators). Despite the existence of multiple examples of AM-related advanced structural optimization methods and multiple examples for AM of bellows structures, there is a lack of knowledge on the design and use of functional multi-material structures in industrial-near applications. This is especially the case for PolyJet™ printing. Thereby, the absence of sufficient knowledge on the achievable performance and sustainability [28] of PolyJet bellows actuators holds back the exploitation of the technological possibilities. The aim of this work is therefore to develop a complex pneumatic-mechanical multi-material component in the context of the development of a PolyJet printable light-weight robot for the DIMAP project. The chosen component is a multi-material light-weight gripper that comprises a geometrically complex bellows actuator and a numerically optimized light-weight structure. By means of the gripper development, structural and material properties are investigated, and suitable design strategies are proposed. With the final design and application to a functional light-weight robot an example is given that may inspire and encourage others for the application of the proposed materials, manufacturing technology and design approaches.

1.6. Design Approach and Concept

The structure of the gripper development is displayed in Figure 1. In a first step a conceptual design (A) is presented that already considers the functional principle of the gripper and adjacent construction in the robot. Moreover, standard parts (bearings, machined parts) are already defined at this stage. The further development is separated into the design of the bellows actuator and the support structure. This separation does not only take place on a component but also on a methodical level as the bellows actuator is (at least for the linear case) optimized via SO and the support structure via TO. Further, the development process of the bellows actuator is divided into a linear (B) and rotary (C) phase. The content of the linear phase was already presented in the conference paper [1] and contains the design and SO of linear PolyJet bellows actuators. Finite elements analysis (FEA) and experiments are conducted to understand the effects of shape and material on the number of endured cycles under repeated loading conditions. The main findings of (B), in particular an initial wall

thickness distribution, the selection of an elastomeric material and a respective strain limit for repeated loading conditions, are translated into the design of a rotary type bellows actuator (C).

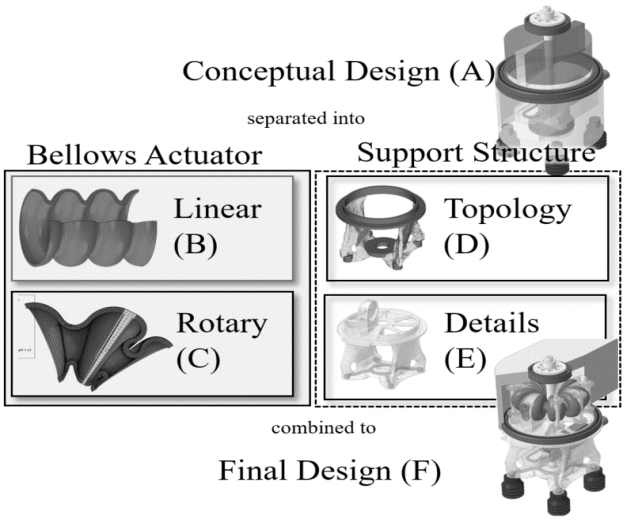


Figure 1: Development approach utilized for the design and development of a PolyJet-printed multi-material light-weight vacuum gripper.

The support structure is investigated separately in (D) and (E). Therefore, the fundamentals of density-based minimum (weighted) compliance TO are developed (D) and applied to the gripper's support structure. The optimization result is smoothened and completed with design features (E). Finally, the rotary bellows actuator and the light-weight structure are combined to the final design (F) and the application of the gripper on a PolyJet printed light-weight robot is demonstrated.

1.7. Conceptual Design

The work described in this paper contributes to the development of a PolyJet-printable vacuum gripper for a pneumatically actuated light-weight robot. When mounted to the robot as shown in Figure 2, the gripper has to fulfill two main functions. First: an object has to be gripped i.e. lifted and carried in a pick and place application. Second: the orientation of the gripped object has to be modifiable i.e. the gripper has to perform a rotatory motion. To drive the design and development process into the direction of an industry-grade performance, a payload of 1 kg and a minimum of 10000 sustainable load cycles are set as requirements. In Figure 2 the conceptual design of the light-weight gripper is displayed. The main components are a rotary bellows actuator on top of the gripper (1), a support structure (2) and a vacuum nozzle (4a) with an adapter (4b). To rotate the gripper, a pressure differential is applied to its two deformable pressure chambers (1a, 1b) that results in a torque about the main axis of the hollow shaft (8). The two antagonistically arranged chambers (1a, 1b) are attached to a fixed (1c) and a movable flange (1d) so that, depending on the external loading condition, a rotatory motion of the gripper relative to the robot (3) can be achieved. In order to grip the payload, a vacuum is created in a vacuum nozzle (Festo VN-05-H-T2, Festo AG & Co. KG, Germany) and guided through the structure to four suction cups (5, Festo ESS-20-CN, Festo AG & Co. KG).

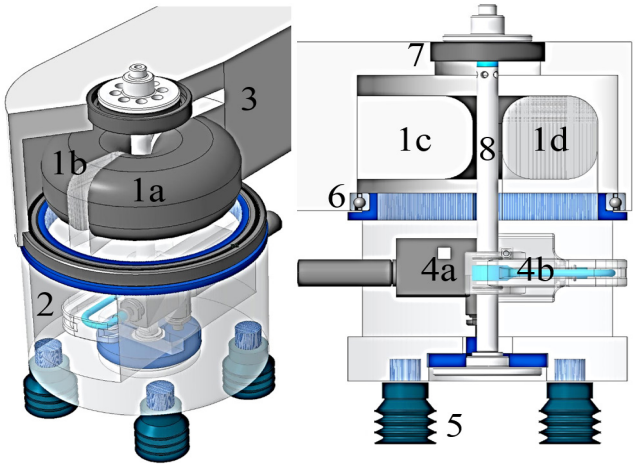


Figure 2: Conceptual design of a PolyJet-printable light-weight gripper with integrated bellows actuator. The main components are the bellows actuator (1), the support structure (2) and a vacuum nozzle (4).

Therefore, compressed air is guided through the hollow shaft (8) into the adapter (4b) and the vacuum nozzle (4a). The hollow shaft (8) is pretensioned against a lower (6) and an upper bearing (7). In order to minimize the contact pressure at the polymeric bearing seats of the attached structural component (3) and keep the weight down, thin section ball bearings (Schaeffler AG, Germany) are chosen. The bottom bearing is a INA CSCA 030 (76,2x88,9x6,35 mm^3, Schaeffler AG) and the top bearing is a CSCAA 010 TN (25x37x7 mm^3, Schaeffler AG).

2. Bellows Actuators

2.1. Linear Bellows Actuators

The behaviour of linear type bellows actuators is investigated first, as their structural behaviour is simpler to analyse by FEA. Linear bellows actuators were designed (Figure 3), comprised of an elastomeric bellows structure and thermosetting flanges.

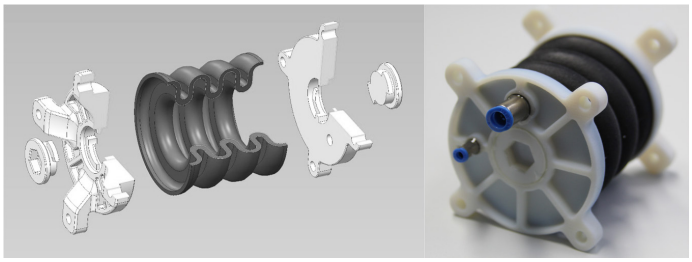


Figure 3: CAD drawing (left) and photograph (right) of a multi-material linear bellows actuator. The bellows structure and flanges are printed in one piece.

The bellows structure and flanges were printed as one monolithic piece and complemented with closing caps and pneumatic connectors (Figure 3). The elastomeric materials TangoBlackPlus™ (TB+) and Agilus30™ (A30) were used. For the flanges VeroWhitePlus™ (VW+) was selected. Actuators with bellows structures made of TB+ were provided by cirp GmbH (Römerstraße 8, 71296 Heimsheim, Germany) and actuators with bellows structures made of A30 by Stratasys® (Haim Holtsman St. 1, 7612401 Rehovot, Israel).

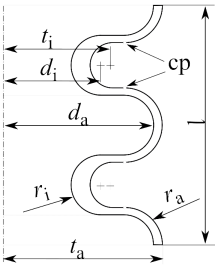
The linear deflection x and "effective force" F_{eff} of a pneumatic actuator depend on multiple factors including the actuator geometry, actuator material and the components of the surrounding pneumatic-mechatronic actuation system. Assuming frictionless guiding and quasi-static conditions, the effective force F_{eff} can be expressed as a function (1) of the pressure force F_{p} and the structural force F_{s} that is caused by the bellows deformation.

$$F_{\text{eff}} = F_{\text{p}} - F_{\text{s}} \tag{1}$$

Thereby, F_{p} can easily be determined ($F_{\text{p}}=(\Delta p)\cdot A_{\text{eff}}$) by multiplying the relative pressure Δp ($\Delta p = p_{\text{i}} - p_{\text{a}}$) and the effective area A_{eff} of the pneumatic chamber. The structural force F_s however, depends on the displacement x and Δp and typically results from an FEA. The curve characteristics of F_{s} (x, Δp) depend on the specific bellows geometry and material. Therefore, a geometrical representation and a material model for elastomeric structures of linear bellows actuators are described below.

2.2. Finite Elements Analysis of Linear Bellows

The geometric model is reduced to the bellows structure because the elastic modulus of the flange material is significantly larger than the modulus of the used elastomers (> 1000 MPa [29], compared to 0.5 MPa [30]). In the FEA described below, the flange is represented by boundary conditions applied to the bellows structure's mesh. For the axisymmetric linear bellows structure (Figure 4), a u-shaped design is chosen that consists of semicircles and parallel lines. The entire bellows structure is defined by 7 design parameters (Figure 4, left). In Figure 4 (right), parameter values (in mm) for an initial (V1) and optimized version (V2) are given and will be referred to in the following.



Param.	Geom. V1	Geom. V2
d_a	30	30
r_i	7.5	5.0
r_a	7.5	6.5
t_a	32.0	32.1
t_i	24.5	21.4
d_i	24.5	20.5
l	52	52

Figure 4: Left: Parameterization of a linear bellows shape with non-constant wall thickness. Control points (cp) are for the evaluation of the distances to adjacent half-waves. Right: Parameter values (Param.) in mm for an initial geometry (Geom. V1) and an optimized geometry (Geom. V2).

The structural force F_{s}, exerted by the deformed bellows structure, originates in the strive of the molecular chains in the elastomeric material to return to their initial configuration. Typically, a strain energy function U is utilized to describe this entropic elasticity in elastomers on a macroscopic scale [31]. In the polynomial form (2) [32], U is expressed as a function of the first and second invariant (I_1, I_2) of the left Cauchy-Green deformation tensor and—in case of compressibility—of the elastic volume strain J_{el} as

$$U = \sum_{i+j=1}^{N} C_{ij}(I_1\text{-}3)^i(I_2\text{-}3)^j + \sum_{i=1}^{N} \frac{1}{D_i}(J_{el}\text{-}1)^{2i}. \tag{2}$$

Thereby, C_{ij} and D_i are constants that are related to the deviatoric and volumetric material behavior respectively. Reducing the general polynomial form (2) to the first order (N=1), the Mooney-Rivlin form for compressible materials [32] is obtained as

$$U = C_{10}(I_1\text{-}3) + C_{01}(I_2\text{-}3) + \frac{1}{D_1}(J_{\text{el}}\text{-}1)^2. \tag{3}$$

Mooney-Rivlin model (3) to uniaxial tensile and compression test data of TB+ with Abaqus' (Dassault Systèmes) internal fitting procedure, the material constants (C_{10}=0.11 MPa, C_{01}=4.52 MPa, D_1 = 2.28 MPa) were determined. A FEA is carried out utilizing the described geometrical and material model. In order to verify the FEA, the effective force F_{eff} for given pressures Δp and deflections x, was measured and compared to simulated results. Figure 5 shows a TB+ bellows actuator mounted to an actuator test bench in three states of enforced displacement. In the testing procedure, the applied pressure difference Δp is controlled by a Festo VPPM pressure control valve (0-2 bar), F_{eff} is measured using a Burster 8523-50 force sensor (+/- 50 N) and the linear displacement x is enforced by a Festo EGSA-50-100 linear axis.

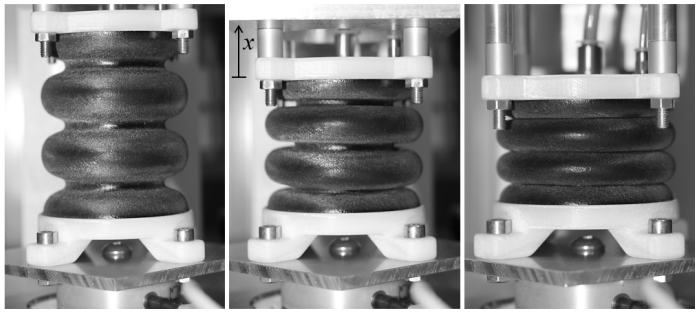


Figure 5: Measuring effective forces F_{eff} of linear bellows actuators for given pressures and enforced displacements. Extension (left), initial (middle) and compression (right) states were tested.

In the experimental procedure, relative pressures Δp were varied in 20 mbar steps from 0 mbar to 140 mbar. Displacements were varied between 20 mm of compression and 30 mm of extension. Noticeably, the effective force was observed to increase for almost 30 s after the enforced displacement states and pressures were reached. Therefore, effective forces at all displacement-pressure combinations were measured after waiting for 30 s. In Figure 6 the experimental and simulated force-pressure-deflection characteristics of a linear bellows actuator are compared. As expected, the effective force F_{eff} exerted by the actuator increases with an increase in relative pressure Δp but generally decreases with an increase in deflection x. For compressions (x < 0), the effective force F_{eff} remains almost constant. Remarkably, none of the interpolated lines covers the full deflection range. Low pressures shift the static force equilibrium (2) and limit the maximum extension. At higher pressures, compression is limited because the waves of the bellows geometry touch adjacent waves ("self-contact", see Figure 5, right). Results from FEA are generally in good agreement with the experimental results. However, Figure 6 indicates, that deviations correlate with high pressures or elongations i.e. increase with increasing strains. For improved significance of the material model, additional stress

states (pure shear, bi-axial) should be considered in the fitting procedure of the constitutive model.

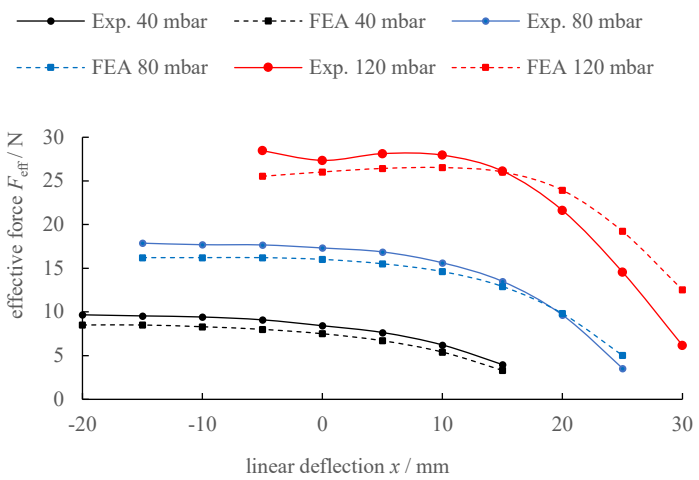


Figure 6: Comparison of the simulated (dashed lines) and experimental results (solid lines) of the effective actuator force F_{eff} as a function of given pressures and forced displacements. Simulations are generally in good accordance with the experiments. Lines are interpolated between measuring points.

2.3. Shape Optimization of Linear Bellows

The underlying concept of a bellows actuator is, that the deformation of a pressurized chamber is utilized to create a linear or rotatory motion. This implies not only the absence of holes in the bounding surface but also leads inevitably to folded structures in which the total actuator displacement is "distributed" into relatively small strains. Consequently, an optimum material lay-out or topology is known a priori (folded structure without holes). Therefore, optimizing a bellows structure can be reduced to the optimization of its shape.

In elastomers that undergo repeated strains [33], [34], material imperfections—also typical for AM materials [28]—cause local strain peaks that can lead to the formation and propagation of microscopic cracks and may eventually result in fatigue failure. In the experiments described above, the bellows structures failed after undergoing repeated deformations i.e. strain. To find an improved bellows design (V2), that reaches similar effective force F_{eff} and deflection x as the initial geometry (V1) but sustains an increased number of load cycles, a numerical optimization routine was developed. Thereby, maximum (logarithmic) principal strain $\varepsilon_{ln,\,max}$ was considered as a fatigue life indicator [35] in the strain objective function

$$Q_{\varepsilon}(x)=(\varepsilon_{ln,\,max}(x)-\varepsilon_{max})^2. \tag{5}$$

During the optimization process, various designs are created. Sets of design parameters in the design vectors x that lead to simulated strains $\varepsilon_{ln,\,max}$ larger than a reference strain ε_{max} result in large values of the objective function. To achieve a required deflection and avoid self-contact, the specific objective functions for the deflection and self-contact $Q_{ld}(x)$ and $Q_{sc}(x)$ are stated analogous to $Q_{\varepsilon}(x)$. The quality of a design is described by the weighted sum of specific objective functions. The resulting multiple criteria objective function is

$$Q(x)=w_{\varepsilon}\cdot Q_{\varepsilon}(x)+w_{ld}\cdot Q_{ld}(x)+w_{sc}\cdot Q_{sc}(x) \tag{6}$$

with w_{ε}, w_{ld} and w_{sc} being weighting factors for the specific objective functions. Therefore, a bellows design is fully described by the design vector

$$x=[r_i \quad r_a \quad t_i \quad t_a \quad d \quad \Delta p]^T. \tag{7}$$

Note that Δp has to be variable and is a component of the design vector because the required effective force F_{eff} is implemented as a hard constraint. Additionally, an integer parameter n_{hw} is defined to quantify the (even) number of half-waves that describe the bellows structure. To account for bounds and secondary constraints (their exhaustive description is beyond the scope of this article), a constraint vector $g(n_{hw}, x)$ is defined. Therefore, the constraint mixed integer bellows optimization problem is

$$\min_{n_{hw}}\left\{\min_{x}\{Q(n_{hw}, x)|g(n_{hw}, x)\leq 0\}\right\}. \tag{8}$$

In order to solve the optimization problem (8), an optimization routine was realized that contains the structural simulation and parameterization described above.

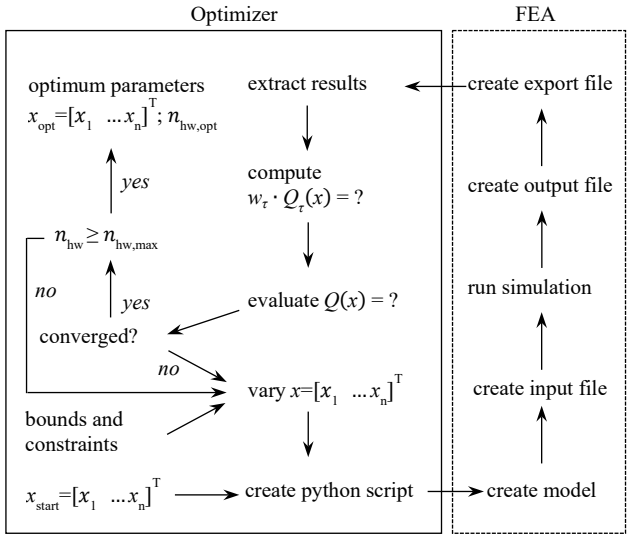


Figure 7: Optimization routine of linear bellows actuators with Matlab (MathWorks) and Abaqus (Dassault Systèmes). The optimization routine finds an optimum bellows shape and corresponding relative pressure concerning a required effective force at defined displacement. Maximum principal strain is minimized in order to increase the fatigue life of the bellows actuator.

Abaqus (Dassault Systèmes) and Matlab (MathWorks) were connected utilizing python scripts as shown in Figure 7. Starting with the initial parameters x_{start} and $n_{hw,start}$ the optimization routine terminates with the output of the optimum parameters x_{opt} and $n_{hw,\,opt}$. The Matlab function *fmincon* with default settings (gradient based optimizer) was chosen for convenient implementation of bounds and secondary constraints. The optimization routine was run with a maximum strain reference of ε_{max} = 0,2 and a force requirement of 12 N at 30 mm linear deflection. In Figure 8 the max. principal strain distribution of the initial (top left) and the optimized bellows geometry (top right) are displayed. The corresponding shape parameters can be found

in Figure 7 (right). The initial geometry (V1) is based on an intuitive design process and comprises four half-waves with constant wall thickness. Applying 140 mbar of relative pressure results in an effective force of 12 N at 30 mm deflection. The corresponding deformation induces a (simulated) maximum principal strain of 65 % that occurs at the inner diameter of the structure (red). The optimized geometry (V2) consists of six half-waves. Wall thickness is non-constant with thickened regions at the inner diameter. Applying 100 mbar at a linear deflection of 30 mm, the effective force is above 12 N. Under these conditions, the (simulated) max. principal strain is 24 % and maximum principal strains are distributed more evenly if compared to the initial geometry (V1) i.e. respective values at the inner and outer diameter are almost equal.

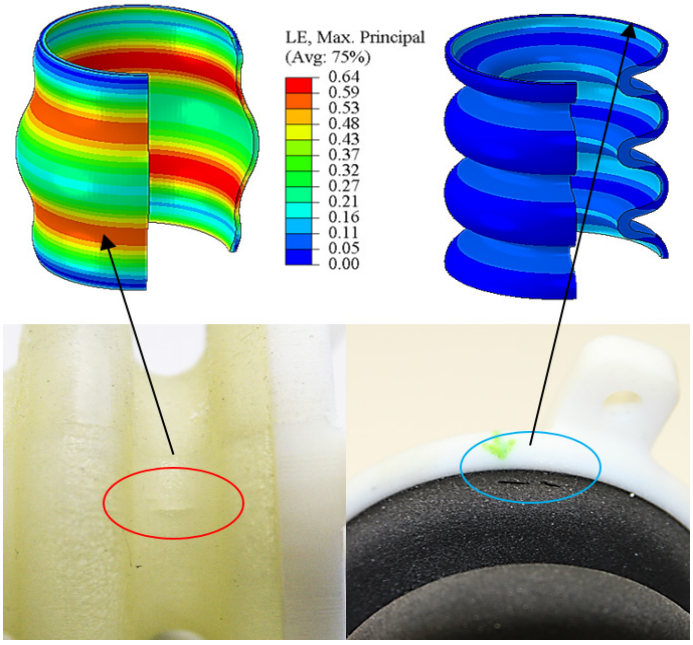


Figure 8: Comparison of the max. principal strain distribution and failures. Cracks in the bellows structure of the initial geometry "V1" (left) are oriented in axial direction occur at the inner diameter. Cracks in the bellows structure of the optimized geometry "V2" (right) are oriented in tangential direction and occur next to the flange. Our observations in the experiments match the locations and are perpendicular to the directions of max. principal strain in the FEA.

Figure 9 shows a comparison of the pressure-force-deflection characteristics of the initial (V1) and optimized (V2) bellows geometry. Noticeably, the effective force of the initial geometry (V1) is significantly more dependent on its deflection. The loss of effective force at increased deflections in Figure 8 is caused by the increased strain in the structure. In the optimized actuator (V2), lower pressures result in similar forces at large displacements. The optimized geometry (V2) satisfies the force constraint of 12 N at 30 mm. Considering the significant reduction of the (simulated) maximum principal strain (from 64 % to 24 %) and fatigue data from literature [28], an increased sustainability of the optimized actuators can be expected.

2.4. Fatigue Testing

In order to examine the hypothesis of an extended fatigue life of the optimized geometry and for comparison of the materials TB+ and A30, endurance runs were performed with the linear bellows actuators. A30 is a recently released PolyJet elastomer with similar hardness range (Shore A 30-35 compared to 26-28 for TB+) as well as a higher elongation at break and tear resistance [36]. Due to superior properties in the data-sheet [36], an increased fatigue life was expected if compared to TB+. The same geometries and material parameters were used for both the TB+ and A30 bellows actuators because no sufficient material data of A30 were available at this time. Therefore, A30 results should be interpreted with care and are presented here for comparison only.

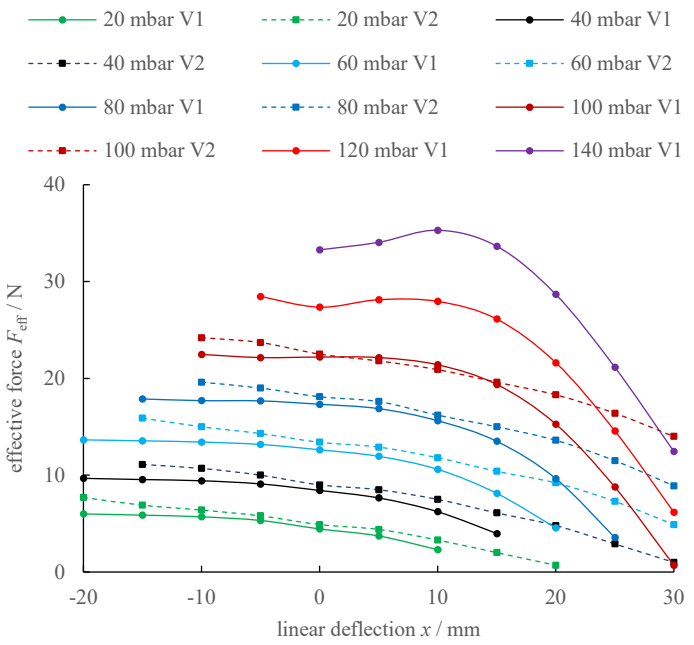


Figure 9: Experimental comparison of the pressure-force-deflection characteristics of the initial (V1, solid lines) and the optimized (V2, dashed lines) bellows geometry. Actuators with the optimized geometry (V2) require significantly less pressure to reach the required forces at 30 mm deflection. Lines are interpolated between measuring points.

In Figure 10, the endurance test bench is displayed. During the test, the right side of the actuator was attached to the test bench (Figure 10). The left side is constrained to horizontal translation by four PTFE-lubricated guiding bolts. The translation was mechanically restrained to 30 mm. In the procedure, a differential pressure of 140 mbar (V1) or 100 mbar (V2) was applied for 30 s. The actuators were vented for another 30 s before the cycle was repeated. Volume flow was measured during the 30 s period to detect possible failure of the bellows structure. The experiment was stopped in case that 2 Nl/min were exceeded.

In Figure 10 the cycles to failure of the tested actuators are displayed. Thereby, cycles to failure range from below 20 (TB+, V1) to more than 30000 (A30, V2). A30 bellows with the initial geometry (V1) endured 143 load cycles in average. With the optimized geometry (V2), A30 bellows endured an average of 24104 load cycles. Despite the relatively small number of specimen (each point in Figure 10 corresponds to a single endurance run), our results indicate that the optimized geometry (V2) sustains significantly more cycles until failure if compared to the initial geometry (V1). Moreover, we conclude that bellows made from A30 can endure more cycles to failure compared to those manufactured from TB+. Noticeably, different geometries lead to two consistent categories (i.e. modes and locations) of

failures. As shown in Figure 8 (bottom), all specimens of the initial geometry (V1) show axial cracks at the inner diameter of the structure (Figure 8, bottom left). In the corresponding FEA, maximum principal strains are oriented perpendicular to the observed cracks and located at the inner diameter. Actuators with the optimized shape (V2) consistently show tangential cracks (i.e. perpendicular to axial strains) near the flanges as shown in Figure 8 (bottom right). To reduce strains at the flange region of the optimized shape (V2) additional shape parameters could be introduced in the optimization routine.

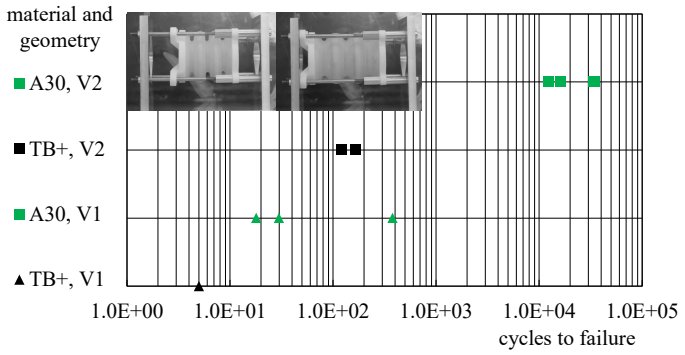


Figure 10: Endurance run with linear bellows actuators load cycles to failure of PolyJet printed linear type bellows actuators. Cycles of pressurization (extension) and ventilation (contraction) were performed until a threshold volume flow was exceeded as an indicator of structural failure. Most cycles to failure were obtained from the combination of Agilus30 (A30) with an optimized bellows shape (V2).

2.5. Rotary Bellows Actuators

The investigations described above pose an intermediate step in the development process of a PolyJet light-weight gripper with integrated elastomeric bellows actuators (number (2) in Figure 8). Three main findings can be formulated as a basis for the initial design of a rotary type bellows actuator.

1. A constant wall thickness leads to extensive strains at the inner diameter of the bellows structure. Thickening this area (as proposed by the shape optimization algorithm), may lead to an increased number of sustainable cycles.
2. A30 should be used instead of TB+ as we observed significantly increased fatigue life for the same geometries and loading conditions.
3. In the performed endurance runs, modes and locations of failure largely corresponded with maximum principal strain in our FEA. Reducing this value to 20 – 25 % delivered acceptable sustainability.

As already displayed in the conceptual gripper model (Figure 2) one rotary actuator comprises two antagonistic (i.e. separate) chambers (1a, 1b). Following this concept and based on the findings of the previous section, an initial geometry was defined in Creo (PTC Inc.) and imported into Abaqus CAE (Dassault Systèmes) for structural analysis and refinement of the bellows shape. The final design of a single chamber for the rotary bellows actuator is displayed in Figure 11. In contrast to existing designs, buckling of the deformable structure is prevented by a roller guiding system and radial forces are supported by the surrounding structure.

As displayed in Figure 12 (top), each chamber is built up from four equal elastomeric segments (1). Between the elastomeric sections, rigid frames (2) are placed that hold rotatable rollers (3). The fixed flange can be attached to the robot link and the movable flange (5) to the gripper's support structure which is pivoted as shown in Figure 2. Under extension or compression, each segment (within one chamber) undergoes the same deformation.

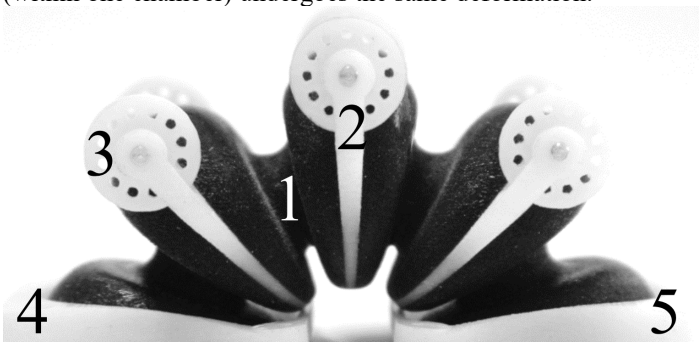


Figure 11: Single chamber for a rotary type bellows actuator. The main components are the soft bellows segments (1), frames (2), rollers (3) and flanges (4,5). The rollers prevent buckling of the bellows structure. The frame-roller combination at the end of each bellows segment subdivide the design into equal design features which facilitates structural analysis and scalability.

As displayed in Figure 12 (top), each chamber is built up from four equal elastomeric segments (1). Between the elastomeric sections, rigid frames (2) are placed that hold rotatable rollers (3). The fixed flange can be attached to the robot link and the movable flange (5) to the gripper's support structure which is pivoted as shown in Figure 2. Under extension or compression, each segment (within one chamber) undergoes the same deformation.

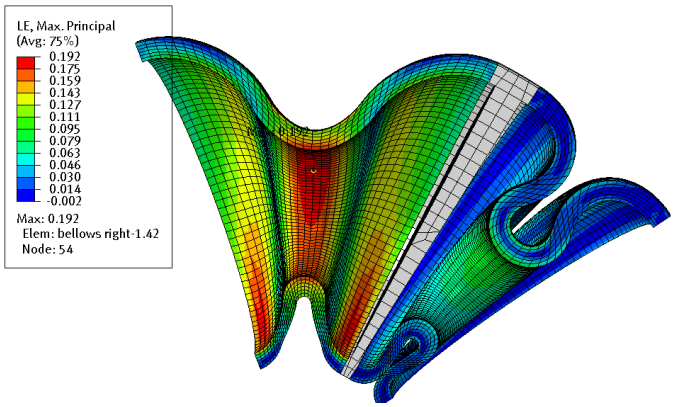


Figure 12: In the printed actuator (top), four identical segments are arranged in series to one of two counteracting chambers. In the corresponding FEA (bottom), two segments are considered because the expansion of one chamber denotes the compression of the antagonistic chamber.

This is due to the serial arrangement within a chamber and the roller guiding system that inhibits buckling. However, the expansion of one chamber denotes the compression of the antagonistic chamber. In the corresponding FE model, the roller guiding system was modelled by constraining the motion of the frame to the rotational degree of freedom (DOF) about the center axis. Hence, the FE model (Figure 12, bottom) was reduced to one half of a bellows segment of one agonist and one antagonist chamber. Nodes at the left and right side are constrained in all DOF. The geometry was meshed so that the bellows structures thickness was represented by 6 (quadratic) elements. For the bellows structure, the hyper-elastic Mooney-Rivlin material model was used as described in paragraph 3 whereas the frame

was modelled as a rigid body. A linearly increasing pressure load was applied to the inner surface of the left bellows segment and left surface of the flange. The analysis type was set to non-linear quasi-static. The occurring (logarithmic, maximum principle) strains were interpreted as a fatigue life indicator. The geometry was then optimized by iteratively adding material at locations of extensive strain and removing material from less strained sections. The final geometry was defined after 5 iterations of re-design in Creo (PTC Inc.) and re-analysis in Abaqus (Dassault Systèmes) and provides a good compromise between the estimated max. rotation angle (95°) at 90 mbar and estimated max. principal strain (19 %). The described FEA does not consider any inertial loads or friction in the roller guiding system and bearings (Figure 2). Therefore, the simulated angular deflection should be interpreted as an upper limit. In initial measurements, a maximum deflection of 50-60° at 90 mbar was observed. The described actuator reaches a theoretical maximum torque of 86 Nmm at 90 mbar and 0° deflections. Apparently, overcoming the frictional torque of the rollers and bearings requires already a considerable fraction of the theoretical torque.

3. Light-Weight Structure

3.1. Density Based Topology Optimization

For the optimization of the bellows shape, a boundary variation approach was applied that is based on an a priori known optimum (or at least very reasonable) material lay-out. However, concerning the light-weight structure of the gripper, no reasonable initial shape can be defined based on intuition or examples. This is due to the complexity of the design space and expected loads and because discontinuities ("holes") in the structure are permissible. The general task for the design of the gripper's support structure is to determine the optimum spatial material distribution within the design space (as shown in Figure 2) that results in a maximum stiffness considering typical or critical loads and boundary conditions. This is exactly what is known today under the term "topology optimization" (TO). In the following, some of the fundamental concepts and equations for density-based minimum (weighted) compliance TO are explained. Detailed descriptions and exemplary Matlab code of density-based TO can be found in [6] and [37].

In typical TO approaches, the objective (and constraint) functions are structural responses from a corresponding FEA. Therefore, an FE model is set up with the discretized geometry being subdivided into design- and non-design regions. The task of the optimizer is to decide for each element of the design region if it should be void or material. As this is clearly an integer problem [8] and the optimization of a large number of discrete variables is computationally time-consuming [38], in popular algorithms ([8], [38]), continuous pseudo densities ρ_e are introduced for each element. Moreover, the stiffness of each element is coupled to its pseudo density that can continuously be varied (9) between 0 (void) and 1 (solid material) by the optimizer. Hence, the design variables of (density based) topology optimization are the continuous pseudo element densities coupled to the elements in the design space.

$$0 < \rho_e \leq 1 \tag{9}$$

However, intermediate densities i.e. values for ρ_e that are not 0 or 1, can generally not be translated into fractions of real-world structures. In the so-called "power-law" or "SIMP" (Solid Isotropic Material with Penalization) approach the stiffness of each element is coupled to its pseudo density in a nonlinear manner (10). $\widetilde{K}_e(\rho)$ is the pseudo or penalized stiffness matrix of an element, ρ_e^p is the elements pseudo density to the power of a penalization factor p (p > 1) and K_e is the initial/non-penalized stiffness matrix of an element.

$$\widetilde{K}_e(\rho) = \rho_e^p K_e \tag{10}$$

By choosing p > 1, intermediate densities become "uneconomical" [6] because the cost (volume or mass) of an element is high, compared to its penalized stiffness. If the individual element stiffness matrices $\widetilde{K}_e(\rho)$ are assembled to a global stiffness matrix $\widetilde{K}$ of the structure, the compliance C can be defined as a reciprocal measure of the structures stiffness as

$$C(\rho) = \frac{1}{2} u^T \widetilde{K} u. \tag{11}$$

In (11), u is the displacement vector. In order to rate the compliance of a structure subjected to multiple load cases i in a single scalar value, the sum of the compliances C_w, of each individual load case C_i is built using weighting factors w_i as shown in (12).

$$C_w(\rho) = \sum_{i=1}^{n} w_i C_i \tag{12}$$

The minimum weighted compliance optimization problem can now be formulated as

$$\min_{\rho} \; C_w \quad \text{subject to} \tag{13}$$

$$\widetilde{K} u = f \tag{14}$$

$$0 < \rho_{min} \leq \rho \leq 1 \tag{15}$$

$$v = V_{d,c} / V_{d,i} \tag{16}$$

with (14) expressing the equilibrium of external (f) and internal forces for a static load case that has to be satisfied for each FEA iteration and (15) defining the value range for the design variables ρ. The volume fraction v constrains the amount of used volume in the design space $V_{d,c}$ in relation to the initial volume of the design space $V_{d,i}$ (16).

3.2. Optimization of Support Structure

For the TO of the gripper's light-weight structure, HyperWorks software was used (Altair Engineering Inc.) with HyperMesh, OptiStruct and HyperView for pre-processing, solving FEA iterations/optimization and post-processing the results respectively. In OptiStruct, the above described SIMP method

was implemented [38]. As the objective function value is computed from the structural responses of an FEA, a suitable FE model was set up first. For an initial analysis, the conceptual geometry (as displayed in Figure 2) was reduced to one fourth and imported to Hypermesh (Figure 13). The geometry was discretized with 43096 CTETRA elements from which 38354 belong to the design space (transparent) and 4742 to the non-design space (blue). As strains were expected to be small and contacts were modelled with rigid body elements, a linear static analysis was set up (this also corresponds to the described approach in section 3.1). Uni-axial tensile testing of VW+ specimen revealed that for small strains the material behavior can be assumed to be linear and isotropic with an elastic modulus of 1809.00 MPa and a Poisson's ratio of 0.42.

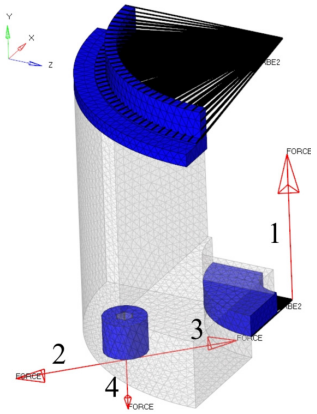


Figure 13: FEA of a light-weight gripper. One fourth of the gripper is modelled including the design space (transparent) and the non-design space (blue).

As shown in Figure 13 loads were applied for pretension of the bearing (force 1: 50.0 N), lifting the payload (force 4: 17.0 N) as well as centrifugal forces resulting from the rotational motion of the whole robot (forces 2 and 3: 35.4 N each). Loads were applied to the independent nodes of 1D RBE2 elements (black). The dependent nodes of the RBE2 elements were attached to the respective nodes in the non-design space with all DOF constrained. It must be pointed out, that the definition of two symmetry planes was made to reduce time in the development (computational time and time for design interpretation). The two planes of symmetry are valid for the load cases "pretension" and "lift" but not for the loads resulting from the rotation of the robotic arm. However, loads in radial direction were considered to drive the design towards a structure that is less sensitive to actual dynamic loads in operation. A TO was performed based on the above described FE model and according to the problem definition ((13)-(16)). The volume fraction was set to v=0.15 (16) and the four loads were weighted equally in the weighted compliance calculation (12). $\rho_{\min}$ was set to 0.01 (default). OptiStruct uses gradient-based optimization algorithms [38] in which the sensitivity of the response (C_w) is calculated with respect to the design variables ρ. In this case, convergence was reached within 29 iterations. In Figure 14 (top left) the optimization result is shown. For the illustration, the obtained geometry is mirrored by the (assumed) planes of symmetry. Noticeably, the main feature of the structure is a straight connection from the base-plate at the bottom to the bearing seat at the top. Smaller Features are the connections from the attachment points of the suction cups to the base plate and bearing seat. This result seems meaningful because the dominant load (load 1 in Figure 13) results in a compression of the structure between the base plate and the bearing seat. The features connecting the attachment points of the suction cups to the rest of the structure are generally in accordance with the directions of applied loads (loads 2 and 3: radial; load 4: vertical).

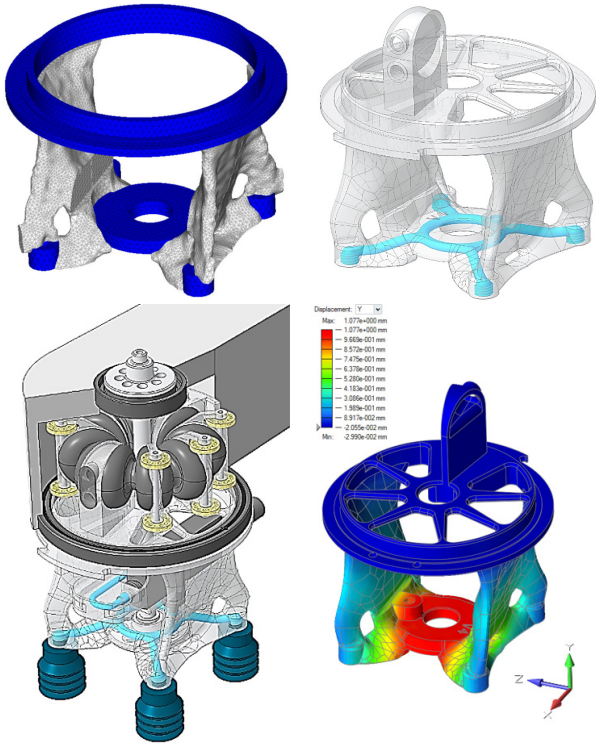


Figure 14: Optimization results and final design. Results of topology optimization (top left) were interpreted with Altair Inspire PolyNurbs (top right). Details were added to achieve the final design (bottom left) and the final structure was re-analyzed (bottom right).

The structure was remodeled using PolyNurbs functionality of Altair Inspire. Moreover, the geometry was re-imported to Creo (PTC Inc.) and features such as threads for the suction cups, internal air guidings (blue colored in Figure 14, top right) and an attachment for the bellows actuator were added. A CAD drawing of the gripper assembly is displayed in Figure 14 (bottom left). A re-analysis was performed (Figure 14, bottom right) and the structure was inspected for strain peaks. When interpreting the results, it must be mentioned, that the elastic modulus of VW+ is significantly rate-dependent which is typical for polymers. The choice of modulus might be considered less important when performing a minimum compliance TO, based on linear static load cases. However, rate-dependency of the modulus in combination with a mixture of long-term static (preload and lift and hold) and dynamic loads (rotatory motion of the robot) bring a certain vagueness to the computation of the weighted compliance (eq. 12) and re-analysis. In further research, the effect of rate-dependent material behavior on the optimization result should be considered.

4. Final Design and Application in Light-Weight Robot

Multiple Grippers were printed by Stratasys ® (Haim Holtsman St. 1, 7612401 Rehovot, Israel) and cirp (cirp GmbH Römerstraße 8, 71296 Heimsheim, Germany). In Figure 15 a prototype of the

developed gripper is displayed. Including bottom and top bearing as well as the adapter with vacuum nozzle, the gripper has a weight of 263 g (Figure 15, left) of which the printed multi-material structure constitutes 102 g.

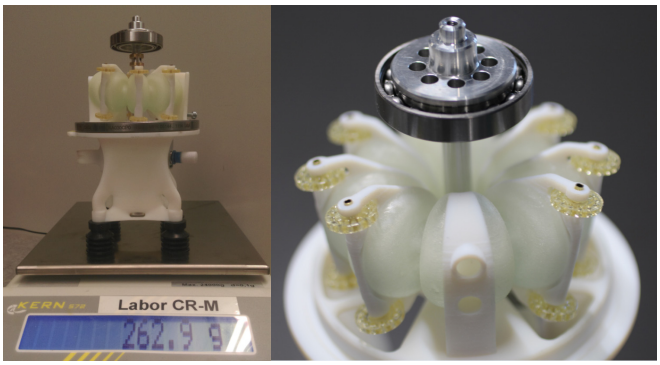


Figure 15: Prototype of a multi-material light-weight gripper with integrated elastomeric bellows actuator. Left: the printed light-weight structure and bellows contribute less than half of the total weight. Right: top-end of the gripper with roller guiding system.

A gripper was mounted to the PolyJet printed light-weight robot that was developed for the DIMAP project (Figure 16). The functioning robot was presented to the public at the formnext2018 faire (19.-22.11.2019, Frankfurt, Germany) at the cirp booth (cirp GmbH Römerstraße 8, 71296 Heimsheim, Germany).

Figure 16: Light-weight gripper attached to the DIMAP SCARA (© Festo).

During the four days, no failure occurred and all the features (bellows actuator, vacuum system, light-weight structure) worked flawlessly. Figure 15 shows the PolyJet printed light-weight robot "DIMAP SCARA" with the attached gripper in a typical pick and place application. Objects of more than 1 kg can be lifted with the described configuration of vacuum nozzle and suction cups.

5. Conclusion

Additively manufactured bellows actuators pose an interesting option for the actuation of future robotic systems as their structural behavior is highly tunable by variations of shape and material. By using state-of-the-art AM technologies such as multi-material PolyJet™ printing, soft actuators can be integrated into stiff and light-weight support structures resulting in highly integrated and geometrically complex components. Though there are many publications that contain AM bellows actuators or suitable light-weight design approaches, the combination of both has hardly been investigated so far. In this article, the simulation-driven design process of a multi-material light-weight vacuum gripper with an integrated bellows actuator is described with the aim to demonstrate possible development approaches and the application in a printed light-weight robot. Based on a conceptual gripper design, investigations were separated into the actuator and support structure development. To understand the structural and fatigue behavior of elastomeric PolyJet bellows, linear type multi-material bellows actuators were designed using VeroWhitePlus™ (VW+) material for the rigid flanges and TangoBlackPlus™ (TB+) for the soft bellows structure. Starting from very few (below 50) sustainable load cycles, optimizing the bellows shape and using Agilus30™ (A30) instead of TB+ material, increased the fatigue life to more than 20.000. Based on these findings, a rotary type actuator was designed that consists of two antagonistic chambers with each chamber being a linear arrangement of equal bellows segments. If compared to existing solutions, the design of the novel guiding system facilitates structural analysis and scalability of the geometry. In order to develop the gripper's support structure, the fundamentals of density-based minimum (weighted) compliance topology optimization were developed and applied to the conceptual gripper design. Considering multiple load cases, an optimum material lay-out was defined. By merging the rotary bellows actuator and the light-weight structure, a fully functional multi-material vacuum gripper was obtained that reaches rotatory deflections of more than 60°, can lift and hold objects of more than 1 kg and has a weight of 263 g of which the printed multi-material structure constitutes 102 g. Finally, the gripper was mounted to a PolyJet printed light-weight robot and its functionality was successfully presented to the public at the formnext2018 faire in Frankfurt, Germany. The described development steps and design solutions can be transferred to other manufacturing technologies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors of this paper thank for the opportunity to be involved in the DIMAP project funded as part of the Horizon 2020 Framework program for research and innovation under the grant agreement no 685937. Moreover, we thank the responsible persons for the organization of the 2018 IEEE international conference on Soft Robotics (RoboSoft) and publication of the conference paper [1]. The software used for the simulation-driven design of the light-weight structure was kindly provided by Altair Engineering Inc..

References

[1] G. Dämmer, S. Gablenz, A. Hildebrandt, and Z. Major "Design and Shape Optimization of PolyJet Bellows Actuators," in *Proc. 2018 IEEE International Conference on Soft Robotics (RoboSoft)*, 2018. https://doi.org/10.1109/ROBOSOFT.2018.8404933

[2] A. Clausen, N. Aage, and O. Sigmund "Exploiting Additive Manufacturing Infill in Topology Optimization for Improved Buckling Load". Engineering, vol. 2, no. 2, pp. 250-257, 2016. https://doi.org/10.1016/J.ENG.2016.02.006

[3] A. M. Mirzendehdel and K. Suresh "A Pareto-optimal approach to multimaterial topology optimization". *Journal of Mechanical Design*, vol. 137, no. 10, pp. 101701-12, 2015. https://doi.org/10.1115/1.4031088

[4] DIN ISO/TS 15066:2017-04 "Robots and robotic devices - Collaborative robots," 2017.

[5] M. Zinn, B. Roth, O. Khatib, and J. K. Salisbury "A new actuation approach for human friendly robot design". The international journal of robotics research, vol. 23, no. 4-5, pp. 379—398, 2004. https://doi.org/10.1177/0278364904042193

[6] M. P. Bendsøe and O. Sigmund, Topology optimization: theory*, methods, and applications.* Springer Science+Business Media, 2013, ISBN 978-3-662-05086-6. https://doi.org/10.1007/978-3-662-05086-6

[7] R. V. Kohn and G. Strang "Optimal design and relaxation of variational problems, I". *Communications on pure and applied mathematics,* vol. 39, no. 1, pp. 113-137, 1986. https://doi.org/10.1002/cpa.3160390107

[8] M. P. Bendsøe "Optimal Shape Design as a material distribution Problem". *Structural and Multidisciplinary Optimization*, vol. 1, no. 4, pp. 193-202, 1989. https://doi.org/10.1007/BF01650949

[9] J. Liu, A. T. Gaynor, S. Chen, Z. Kang, K. Suresh, A. Takezawa, L. Li, J. Kato, J. Tang, W. Jinyuan et al. "Current and future trends in topology optimization for additive manufacturing". *Structural and Multidisciplinary Optimization*, vol. 57, no. 6, pp. 2457-2483, 2018. https://doi.org/10.1007/s00158-018-1994-3

[10] J. Liu, and Y. Ma "A survey of manufacturing oriented topology optimization methods". Advances in Engineering Software, vol. 100, pp. 161-175, 2016. https://doi.org/10.1016/j.advengsoft.2016.07.017

[11] T. Zegard and G. H. Paulino "Bridging topology optimization and additive manufacturing". Structural and Multidisciplinary Optimization, vol. 53, no. 1, pp. 175-192, 2016. https://doi.org/doi:10.1007/s00158-015-1274-4

[12] A. T. Gaynor, N. A. Meisel, C. B. Williams, and J. K. Guest "Multiple-material topology optimization of compliant mechanisms created via PolyJet three-dimensional printing", Journal of Manufacturing Science and Engineering, vol. 136, no. 6, pp. 0610151-10, 2014. https://doi.org/10.1115/1.4028439

[13] S. Junk, B. Klerch, L. Nasdala, and U. Hochberg "Topology optimization for additive manufacturing using a component of a humanoid robot," in Proc. 28th CIRP Design Conference, pp. 102-107, 2018. https://doi.org/10.1016/j.procir.2018.03.270

[14] P. W. Christensen and A. Klarbring, *An Introduction to Structural Optimization*. Springer Science+Business Media, 2008, ISBN 978-1-4020-8666-3.

[15] T. Lens, J. Kunz, O. Von Stryk, C. Trommer, and A. Karguth "Biorob-arm: A quickly deployable and intrinsically safe, light-weight robot arm for service robotics applications," in *Proc. 41st International Symposium on Robotics (ISR)* and *6th German Conference on Robotics (ROBOTIK)*, pp. 905-910, 2010.

[16] M. Grebenstein, A. Albu-Schäffer, T. Bahls, M. Chalon, O. Eiberger, W. Friedl, R. Gruber, S. Haddadin, U. Hagn, R. Haslinger et al. "The DLR hand arm system," in *Proc. 2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3175-3182, 2011. https://doi.org/10.1109/ICRA.2011.5980371

[17] B. Vanderborght, A. Albu-Schäffer, A. Antonio, E. Burdet, D. G. Caldwell, R. Carloni, M. Catalano, O. Eiberger, W. Friedl, G. Ganesh et al. "Variable impedance actuators: A review". *Robotics and autonomous systems*, vol. 61, no. 12, pp. 1601—1614, 2013. https://doi.org/doi:10.1016/j.robot.2013.06.009

[18] A. J. Veale and S. Q. Xie "Towards compliant and wearable robotic orthoses: A review of current and emerging actuator technologies". *Medical engineering & physics*, vol. 38, no. 4, pp. 317-325, 2016. https://doi.org/doi:10.1016/j.medengphy.2016.01.010

[19] D. Baiden and O. Ivlev "Independent torque and stiffness adjustment of a pneumatic direct rotary soft-actuator for adaptable human-robot-interaction," in *Proc. 23rd International Conference on Robotics in Alpe-Adria-Danube Region (RAAD)*, September 2014. https://doi.org/10.1109/RAAD.2014.7002257

[20] R. MacCurdy, R. Katzschmann, K. Youbin, and D. Rus "Printable Hydraulics: A Method for Fabricating Robots by 3D Co-Printing Solids and Liquids," in *Proc. 2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016. https://doi.org/10.1109/ICRA.2016.7487576

[21] A. Grzesiak, R. Becker, and A. Verl "The bionic handling assistant: a success story of additive manufacturing". *Assembly Automation*, vol. 31, no. 4, pp. 329—333, 2011. http://dx.doi.org/10.1108/01445151111172907

[22] D. Drotman, S. Jadhav, M. Karimi, M. T. Tolley et al. "3D printed soft actuators for a legged robot capable of navigating unstructured terrain," in *Proc. 2017 IEEE International Conference on Robotics and Automation,* ICRA, pp. 5532-5538, 2017. http://dx.doi.org/10.1109/ICRA.2017.7989652

[23] B. N. Peele, T. J. Wallin, H. Zhao, R. F. Shepherd "3D printing antagonistic systems of artificial muscle using projection stereolithography". *Bioinspiration & Biomimetics*, vol. 10, no. 5, 2015. http://dx.doi.org/10.1088/1748-3190/10/5/055003

[24] B. Mosadegh, P. Polygerinos, C. Keplinger, S. Wennstedt, R. F. Shepherd, U. Gupta, J. Shim, K. Bertoldi, C. J. Walsh, G. M. Whitesides "Pneumatic networks for soft robotics that actuate rapidly". *Advanced Functional Materials*, vol. 24, no. 15, pp. 2163-2170, 2014. http://dx.doi.org/10.1002/adfm.201303288

[25] I. Gaiser, A. Andres, G. Bretthauer, H. Breitwieser, O. Ivlev, R. Wiegand, S. Schulz "Compliant robotics and automation with flexible fluidic actuators and inflatable structures," *INTECH Open Access Publisher*, 2012. http://dx.doi.org/10.5772/51866

[26] A. Zolfagharian, A. Z. Kouzani, S. Y. Khoo, A. A. A. Moghadam, I. Gibson, A. Kaynak "Evolution of 3D printed soft actuators". *Sensors and Actuators*, vol. 250, pp. 258-272, 2016. https://doi.org/10.1016/j.sna.2016.09.028

[27] P. Polygerinos, N. Correll, S. A. Morin, B. Mosadegh, C.D. Onal, K. Petersen, M. Cianchetti, M. T. Tolley, R. F. Shepherd "Soft Robotics: Review of Fluid-Driven Intrinsically Soft Devices; Manufacturing, Sensing, Control, and Applications in Human-Robot Interaction". *Advanced Engineering Materials*, 2017. http://dx.doi.org/10.1002/adem.201700016

[28] J. P. Moore and C. B. Williams "Fatigue properties of parts printed by PolyJet material jetting". *Rapid Prototyping Journal*, vol. 21, no. 6, pp. 675-685, 2015. http://dx.doi.org/10.1108/RPJ-03-2014-0031

[29] O. Sheikhnejad, F. Hiptmair, and Z. Major "Finite element simulation and mechanical modelling of nanoparticles reinforced composite," in *Proc. 6th International Conference on Additive Technologies*, 2016.

[30] M. Reiter and Z. Major "A combined experimental and simulation approach for modelling the mechanical behaviour of heterogeneous materials using rapid prototyped microcells". *Virtual and Physical Prototyping*, vol. 6, no. 2, pp. 111-120, 2011. http://dx.doi.org/10.1080/17452759.2011.586949

[31] R. W. Ogden, *Non-linear elastic deformations*. Dover Publications, Inc., Mineola, 1997, ISBN: 9780486696485.

[32] Dassault Systemes Simulia Corp "Abaqus Theory Guide. Version 6.14,", 2014.

[33] A.N. Gent, P.B. Lindley, and A.G. Thomas "Cut growth and fatigue of rubbers. I. The relationship between cut growth and fatigue". *Journal of Applied Polymer Science*, vol. 8, no. 1, pp. 455-466, 1964. https://doi.org/10.1002/app.1964.070080129

[34] G.J. Lake and P.B. Lindley "Cut growth and fatigue of rubbers. II. Experiments on a noncrystallizing rubber". *Journal of Applied Polymer Science*, vol. 8, no. 2, pp. 707-721, 1964. https://doi.org/10.1002/app.1964.070080212

[35] Y. Zhou, "Fatigue Properties Of Magnetorheological Elastomers And The Design Of Interfacial Layers To Improve Fatigue Life." Ph.D. thesis, Dublin Institute of Technology, Ireland, 2016. https://doi.org/doi:10.21427/D75315

[36] Stratasys Ltd, "PolyJet Materials DataSheet", [Online]. Available: http://usglobalimages.stratasys.com/Main/Files/Material_Spec_Sheets/MSS_PJ_PJMaterialsDataSheet.pdf?v=635785205440671440. [Accessed: Feb. 26. 2018].

[37] O. Sigmund "A 99 line topology optimization code written in Matlab". *Structural and Multidisciplinary Optimization*, vol. 21, no. 2, pp. 120-127, 2001. https://doi.org/10.1007/s001580050176

[38] OptiStruct, Altair "OptiStruct 11.0 user manual", Altair Engineering Inc, Troy, MI Google Scholar, 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 27-39 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Practical PIR-based Scheme for Discovering Nearby Places for Smartphone Applications

Maryam Hezaveh*, Carlisle Adams

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Article history:
Received: 10 December, 2018
Accepted: 14 January, 2019
Online : 20 January, 2019

Keywords:
Private Information Retrieval
Privacy
Smartphone applications
Location-Based Services

ABSTRACT

We present a privacy-preserving approach for discovering nearby places of interest to Alice. In this approach, the proposed protocol allows Alice to learn whether there is any place that she is looking for near her. However, the location-based service (LBS) that tries to help Alice to find nearby places does not learn Alice's location. Alice can send a request to the LBS database to retrieve nearby places of interest (POIs) without the database becoming aware of what Alice fetched by using private information retrieval (PIR). The common criticism of previous PIR approaches is that they are not practical for smartphones with limited processing power, memory, and wireless bandwidth due to the computational overhead. Therefore, the main focus of this work is to propose a scheme to reduce the computation cost on the client-side to make PIR appropriate and practical for the smartphone environments, and then apply the proposed PIR to LBS applications. We have implemented our protocol in Percy++ to evaluate its performance over a commercial-grade database of POIs. Our implementation results demonstrate that our approach has faster decode and retrieve time for the nearby POIs on smartphones compared with current similar work.

1. Introduction

Location-based services (LBS) are information services that offer various types of applications based on the location of the user, such as identifying a location of a person or object or place, weather service, parcel and vehicle tracking, etc. LBS retrieves the location of the user from the user's mobile phone via global positioning system (GPS), cell tower triangulation, or wireless local area network (WLAN). LBSs might be helpful to mobile users for safety services. For example, it is beneficial if emergency services can find information about the location details of a user who is in danger. However, users may not be aware that their location information might be shared with other third parties and could be misused, and it could be diverted as a tracking tool. Therefore, the main goal for research communities in this field is to protect the user's location while they are using LBSs.

During the last two decades, privacy-preserving protocols for location-based services have been introduced based on non-cryptographic and cryptographic approaches. Non-cryptographic approaches use trusted third parties to maintain the user's privacy, such as "dummy locations," "K-anonymity" and "cloaking" approaches. Reviews of these approaches and their drawbacks can be found in [1, 2, 3, 4, 5]. The main disadvantage of these approaches is that the user has to trust and send her information to the trusted third party, which as we mentioned is unacceptable in LBS applications.

Our main goal in this paper is to help LBS users search on their smartphone for particular places of interest (POIs) while keeping their location private from the location-based service provider (LBSP). For example, the user sends her request to the LBS application to find a nearby restaurant, gas station, or ATM. This location could be the exact location of the user or a location to which she wants to travel in the future. These types of applications are especially useful for people who travel or have moved to a new city. A wide range of LBS nearby places applications have been released recently, such as Facebook Nearby Places, AroundMe, NearBy Places, Yelp, FourSqure and Places to help users to identify their nearby locations quickly.

Modern multi-core mobile devices have high-performing processors that are appropriate for cryptographic tasks which can enable location privacy to the LBS applications. Unfortunately, the processing units frequently consume a significant amount of energy, which causes a reduction in the battery life of smartphones. In addition, smartphones devices often have limited bandwidth and memory [6]. Therefore, downloading an entire database for finding POIs around Canada and the U.S., which easily can include more

*Maryam Hezaveh, University of Ottawa, mheza028@uottawa.ca

than ten million entries with respect to a typical commercial POI database (see Appendix A) [7] and require 3 to 4 GB of data storage, is obviously not practical. Furthermore, updating results periodically to make sure they are accurate enough, bandwidth limitation and data usage limitation of smartphones are other important factors which we should keep in mind when we are offering a cryptographic solution for LBS applications with respect to the privacy of the mobile user's location.

1.1. Motivation and Threat Model

Enormous enthusiasm for geographical referencing of individual information is apparent on the web these days. The majority of people currently use smartphones with many complex sensors closely connected to their daily activities. Most of these smartphones have a high-precision localization sensor such as a GPS receiver. GPS devices allow people to tag photographs and occasions and to track their mobility. Moreover, the number of sensors in our environment which interact with smartphones has increased. Although most people like the convenience of using these personal communication devices, there is an inherent trade-off between convenience and privacy. Clients might not be completely aware of exactly how the data about their location is utilized and by whom and what information about their location is being gathered, and subsequently, clients can disregard the potential risk that can happen by using their location information.

Location-aware capabilities allow the service providers to offer different types of application to their users such as the ability to share their location with their friends and other users and to geo-reference their posts. In this way, clients can utilize the location identifier to search and browse for different resources. An essential key for providing these services is to gather real-time location information on clients and additionally, other logical data including client relationships, activities and client provided content perhaps during long intervals of time. Specifically, in nearby location applications, service providers are not only able to collect clients' location information, but are also able to gather personal information by offering clients to write their experience and opinions with regard to reviews and tips on the visited place. Subsequently, clients' historical location data can be identified with relevant and semantic data freely accessible online and can be utilized to discover individual and sensitive data about clients and to develop comprehensive client profiles. The user activities, relationships, interests and mobility patterns could be extracted from these profiles. Although these location-based profiles may be considered helpful to improve and personalize the quality of applications for the clients, they can potentially be utilized for unwanted purposes and can cause different levels of privacy threats. Users' mobility tracks are not only a collection of locations on a map. The content of these tracks includes the users' interests, activities, habits, and relationships. It may also disclose users' private information and secrets. It can expose the users to undesirable commercial and spams, or even threat of physical harm. All of these imply that the negative side-effects of lacking location privacy have increased.

The main aim of this paper is to protect the users' location privacy against a passive adversary, active adversary and malicious service providers while they are using LBSs. We consider the following threats in our architecture:

Passive adversary. A malicious external observer or a malicious LBSP who has access to the data that passes between the user and the database on the communication channel but cannot change the data.

Active adversary. A malicious external observer who has access to data that passes between the user and the database on the communication channel and can insert, modify or delete data.

Malicious Service provider. A malicious server refers to an LBSP that tries to modify, delete or insert new messages in response to the user.

Users should have the privilege of controlling the amount of information (about their location) that is revealed and shared with others. This can be achieved in different ways such as users have a right to choose not to share their location information to untrusted applications, legislating privacy policies to force organizations and service providers to protect their users' location privacy, and designing a system in a privacy-preserving manner so it does not disclose users' location information to others.

1.2. Our Contributions and Assumptions

This current paper is an extension of a paper originally presented in [8]. We first explain our proposed block-based PIR scheme for smartphone applications [8] and then as an extension we apply our proposed PIR scheme to the LBS application. Our privacy-preserving protocol for LBS helps Alice to search for specific nearby POIs on her smartphone by sending a query to the location based service provider (LBSP) over a wireless network. In this scenario, the proposed protocol allows Alice to learn whether there is any place that she is looking for near her. However, the location-based service (LBS) that tries to help Alice to find nearby places does not learn Alice's location. Alice can send a request to the LBS database to retrieve nearby places of interest (POIs) on her smartphone without the database becoming aware of what Alice fetched by using our practical PIR scheme. The LBS server retrieves the query from the database, and returns the results to Alice containing the specific requested POIs type found in the requested location. In order to achieve this, our protocol must fulfill all of the following requirements, as also required in [9]:

1. The LBS server must not learn the exact location of the user. It might only identify a area that is large enough to satisfy the user's privacy in terms of area and the number of POIs it contains.
2. The proposed protocol must have no third parties between the user and the server.
3. The implementation must be computationally practical for resource-constrained hardware such as a smartphone.

4. The proposed approach cannot depend on trusted hardware that does not generally exist on a commercial smartphone.

Our cryptographic approach is based on private information retrieval (PIR) for secure LBS applications that identify nearby places. PIR allows the user to fetch her required information from the database without leaking which information is fetched [10]. The POI database is labeled by the location of POIs; therefore, the LBS server is able to retrieve the POIs depending on the user's location of interest in the requested query. PIR solves most of the previous problems associated with non-cryptographic approaches in LBS. PIR approaches do not have the privacy vulnerabilities of k-anonymity or cloaking, such as single point of attack of their anonymizer or server which tries to help them to apply k-anonymity or generate an obfuscation area. As a result, the information of the user location remains private and secure from all kind of the passive adversary, active adversary and service provider by using PIR approaches.

During the past two decades, various types of PIR-based approaches have been introduced. The common criticism of previous PIR approaches is that they are not practical for smartphones with limited processing power, memory, and wireless bandwidth due to the computational overhead [11, 12]. We ensure that the proposed cryptographic PIR approach is practical for smartphone applications. Based on [13], there are five main time elements that influence the speed of the PIR query:

1. the amount of time that it takes for the client to create a query which has to be private.
2. the amount of communication time that it takes to send the query to the server(s).
3. the amount of time for the server(s) to apply the query to the database.
4. the amount of communication time that it takes to send the response from the server(s) to the client.
5. the amount of time that it takes for the client to decode the response(s) and retrieve the results.

Our approach expands [9] idea of applying a cloaking area to reduce these five factors. Moreover, our approach reduces the amount of time required for the client to process the response and retrieve the results of decoding on the smartphone by approximately 50% compared to [9] by applying the POI types idea to block-based PIR. Reducing the decode time is valuable in our application to satisfy the fifth requirement, so that it can be used on modern smartphones' hardware. The processing cost on the server side is similar to [9], to preserve the privacy of the user's location. Our proposed protocol can be made to support all types of block-based PIR schemes.

In our proposed approach, the identity of the user is not hidden from the service provider, as the results have to be returned to the user. However, if the user wants to keep her identity hidden from LBS, she can use an onion routing technique, such as Tor [14]. Note that keeping the user's location private has priority in an LBS application over keeping the user's identity hidden from LBS, because if LBS knows the user's location, it is quite easy to identify the user. We should mention that a mobile communications operator is constantly aware of the location of the user based on the cell tower. Therefore, we assume that this operator does not collude with the LBSP.

1.3. Organization of This Work

The rest of this paper is structured as follows. Section 2 presents an overview of previous work regarding PIR schemes and LBS schemes. Section 3 describes the details of our PIR scheme. Section 4 explains the details of our privacy-preserving protocol for LBS. The threat model and the security analysis of our proposed protocol are discussed in Section 5. Section 6 gives an overview of our implementation and compares it with previous work. The limitations of our proposed protocol are discussed in Section 7, and finally Section 8 concludes our paper.

2. Related Work

For greater understanding, we first review the definition of PIR and give a brief overview of different types of PIR schemes. Then, we provide a review of PIR-based approaches for the users' location privacy in LBS applications.

2.1 Review of Private Information Retrieval (PIR)

These days, users are increasingly aware of the privacy requirements of their data in their online activities. But is it actually possible to keep the user's query contents private while she issues a request to online applications? The first answer that comes to your mind when you think about this problem is that the user can send her request to the online application via Tor and communicate over the Tor network [14]. Here, the server has no clue who sent the request for the data; however, in order to fetch the requested data from the database, the server has to be able to access the content of query. Therefore, Tor is not a good option to solve our problem. The main problem that we need to solve is to let a user to send her query to the database without sharing what she searched for. In this scenario, we are trying to protect the content of the query, rather than the identity of the user. Private information retrieval (PIR) is a cryptographic technique that solves the matter of permitting the user to query a database while the content of the user's query is hidden from the database. The need for PIR schemes has been demonstrated in real online activities, such as location-based services, social networks, online research, etc. [9].

In 1995, [10] first introduced the problem of Private Information Retrieval (PIR). Looking at the trivial solution [10] of transferring the entire database to the user to be locally queried, highlights interesting properties. First, it delivers perfect privacy. Second, no information about query or response is leaked, since neither of these are sent across the wire. On the other hand, this approach yields high communication overhead: the size of the whole database. Goldberg [15] presented three important requirements for PIR: privacy, non-triviality, and correctness. For privacy, the database should learn neither the query input nor the

database block retrieved. For non-triviality, communication cost between the client and the server should be less the trivial limit of $O(n)$, where n is the number of bits in the database as seen above. For correctness, the received data from the database must satisfy the user's query. Another requirement which is not considered in most of the previous work for PIR is implementation efficiency. Most of the previous work tried to reduce the communication overhead rather than the computational overhead [16, 17]. This inattention to the computational complexity has caused the introduction of PIR schemes that are not practical for resource-constrained hardware, such as smartphones.

In [10], the author defined the first non-trivial PIR scheme. In 2004, Gasarch [18] described it simply as follows.

Definition 2.1. A one-round k-databases Private Information Retrieval (PIR) scheme with x$\in \{0,1\}^n$ is defined as follows [10, 18].

1. A user wants to find x_i. There exists k databases which all have the same copy of $x = x_1 \dots x_n$. The DBs do not collude with each other.
2. The user flips coins and the combination of the coin flips and i, produces query strings $q_1 \dots q_k$. She sends the query, q_j, to database DB_j.
3. For all j queries, where $1 \leq j \leq k$, DB_j returns an answer string $ANS_j(q_j)$.
4. The user computes x_i using the value of the $ANS_j(q_j)$, the coin flips, and i.

The cost of the defined PIR scheme is $\sum_{j=1}^{k} |q_j| + |ANS_j(q_j)|$.

Computational PIR (CPIR): The first type of PIR protocols assumes that the adversary and the server(s) have access to limited computational capability to guaranty the user's privacy. Therefore, to breach the security of these protocols, the adversary has to solve a problem which is hard to solve with its limited computational capability. This kind of assumption is usual for cryptography, security, and privacy schemes.

In 1995, [10] proved that it is impossible to have a single-database PIR in the information theoretic security sense. In 1997, [19, 20] proposed the first CPIR to prove that the communication complexity of PIR can be reduced if we want to achieve computational privacy, and we are not willing to achieve information theoretic privacy. In the same year, Kushilevitz and Ostrovsky [21] presented a CPIR protocol which has the same assumption for the computational capability for the adversary, but it uses a single server. Their protocol was the first single-server CPIR. It is based on the Quadratic Residuosity problem that is considered to be difficult to solve. The main advantage of single-server CPIR protocols is that by using the CPIR recursively, the communication complexity of PIR can be improved. Later, different types of single-server CPIR were proposed which tried to reduce the communication cost of PIR, for example, φ-hiding problem [22, 23], the presence of one-way trapdoor permutations [24], Pailler homomorphic encryption [25], and the Hidden Lattice problem [16].

In [17], the author proved that none of the previous CPIR schemes were practical, given certain realistic assumptions at the time. However, in 2016, [26] introduced XPIR. They showed that by using lattice-based cryptography, CPIR is of practical value and the conclusion of [17] is no longer valid.

Information Theoretic PIR (IT-PIR): In Information theoretic privacy even if an adversary has unlimited computational capability, he cannot compromise the privacy of the user. In 1995, [10] showed that any single-server IT-PIR scheme must have communication cost at least that of the trivial protocol. Therefore, IT-PIR protocols assume that if you have $k \geq 2$ non-cooperating servers, and each of these servers has a copy of the database, then there exist PIR schemes which achieve complete information theoretic security. Following [10], different types of IT-PIR were proposed which tried to improve [10], such as [9, 15, 27, 28].

By using the idea of multiple servers, we improved the robustness of the PIR, but this can affect privacy if there exists non-responsive servers or/and malicious servers [15, 27]. To handle this issue, Goldberg [15] proposed the privacy threshold in which the total number of the servers must be greater than the privacy threshold. As a result, in order to set a privacy threshold, we need to provide extra responding servers.

Trusted Hardware PIR: The trusted hardware-based PIR is first introduced by [29] in 2006. The trusted hardware-based PIR uses the idea of a tamper-resistant CPU, which is connected to the server and is trusted by the user. The user sends her query to this CPU, where her query is hidden from the server. In this scenario, the CPU is the one who is responsible to fetch the requested information from the database and sends back the results to the user. These types of PIR achieve the low computation and communication costs, but the trusted hardware PIR architecture is secure only if the user can trust the hardware.

Hybrid PIR: In [13], the researcher proposed a hybrid PIR that was a combination of CPIR and IT-PIR to reduce communication costs. Their goal was to combine the positive features of CPIR and IT-PIR to reduce the negative features of each. To achieve a lower bound for both computation and communication costs, they merged the recursion property of CPIR (single-server) approaches and the low computation and communication complexity property of IT-PIR (multiple-server) approaches.

2.2 Review of the PIR-based scheme for Nearby Places

One of the motivations for developing useful and practical PIR schemes is to protect the users' private information while they are using mobile devices with positioning capabilities. In a stationary desktop scenario, when a user tries to query the database or the remote server, the primary concern is leaking information about the query's content. However, in an LBS scenario, when a user queries the LBS server, her location is also revealed to the LBS server. Here, the problem with location privacy is preserving the privacy of the user's real location when she is using the LBS while providing the most precise and acceptable response.

Many of the previous problems of privacy preserving protocols for LBS that we encountered were solved by introducing PIR-based LBS protocols. The idea is to let the user send a query to the LBS server without disclosing her actual location by the PIR scheme. This query typically consists of POIs, which includes a description of the POI and its geographic location.

Most of the existing works which tried to apply PIR to location-based services were based on secure hardware, with a secure coprocessor at the LBS server [3, 5, 30, 31]. The idea of using the secure hardware-based PIR in LBS was first proposed by Hengartner [3]. This hardware performs the trusted computing to hide the user's location from the LBSP. Recent work regarding secure hardware PIR was proposed by [30]. Their PIR technique was similar to [31], however it offered better efficiency, and it was more practical for large datasets. All proposed solutions for secure-hardware PIR claim that the trusted hardware-based PIR method is the only practical PIR scheme [30, 31]. The main disadvantage of all secure hardware PIR schemes is that the proposed architectures are secure only if the user can trust the hardware.

The common criticism of other PIR approaches for location privacy is that the computational overhead is not acceptable and practical for resource-constrained hardware such as a smartphone [11, 12]. In 2008, In [1], the author proposed the first PIR-based approach for location privacy, without using a third party. Their proposed protocol used the idea of the trade-off between efficiency and privacy as defined in [32]. In [1], the researcher proposed a single PIR request for each query approach. In their approach, all queries were indistinguishable, and it was able to achieve strong location privacy. Their proposed protocol included two steps to protect the user's query and information about her location. In the first step, the server and the user engaged in a protocol, which is based on Paillier encryption [33], to determine the index of the user's location cell, without releasing the location to the server. The user uses PIR to retrieve the query results for the target cell in the second step. The advantages of the Ghinita protocol are the nondisclosure of location information and its security for both mobile and stationary users against correlation attacks.

In [5], the author described three drawbacks to [1] protocol. First, it focuses on the nearest neighbor queries. Second, it scans the entire database linearly for each query. Third, it has a high communication complexity. Additionally, the protocol is secure if the privacy of the user is a concern and LBS is not able to learn the user's query, but it is not symmetric for LBS's database privacy since the user can infer the data that are in the same column as her query.

Later, in [9], the author proposed a hybrid solution combining PIR and cloaking to protect the user's privacy without using trusted computing. Their idea of using cloaking reduces the computational cost of PIR and makes it more practical. The user's location privacy relies on the size of the cloaking area. Their PIR approach supports all types of PIR schemes (block-based). Our proposed PIR protocol expands on [9] idea. However, we focus on reducing the computational complexity on the client side. We explain our proposed protocol in detail in the next section.

3. The Proposed PIR Scheme

Here we present our block-based PIR for location privacy in mobile phone applications. Our solution uses partial queries [10] to reduce communication and computation complexity. Moreover, we structure the database to optimize client computations. This has benefit in our mobile scenario in which the clients (possibly smartphones or IoT sensors) have constrained computational power. In our approach, the user retrieves the exact category of the data, which saves on data processing on the resulting sets. These savings on result set size in turn impact any decode, decrypt, or homomorphic operations which must occur to obtain a result. As these are cryptographic operations, the benefit in result set size reduction is material. Note that our approach is suitable for all applications that need to protect users' privacy while they are searching for data in a database (it is not restricted to just LBS applications).

3.1. Preliminaries

Our proposed protocol can be made to support all types of block-based PIR schemes. We illustrate its usage using multiple server IT-PIR [15] and Shamir secret sharing [34]. As such, the user's query is split into l shares which are then transferred to k servers. This results in communications and computation benefit which we analyze in Section 6. The protocol is robust to byzantine situations in which servers (either malicious or in a service degradation scenario) may fail to respond or may respond with information containing errors.

Our approach to reduction of client computation cost uses the idea of trading off privacy for better performance [9]. In [9], the level of desired-privacy is adjustable and is proportionally related to the number of data items that the database PIR server must process to respond to the client. We extend and improve on this approach in three ways.

First, we divide the database into classes, with each class categorized based on a sub-type of data to be queried. The server returns exactly the subset of the database which pertains to the queried category. By reducing result set size, the client benefits in a number of ways. It is no longer necessary to filter the response data. In addition, the aggregate cost of cryptographic operations, such as decryption or homomorphic computation, is reduced.

Second, If a sub-type has a higher amount of data, a data traffic cost will be higher because of the result size and it will cause a slower response time. On the other hand, if in another sub-type the amount of data is extremely low, it will minimize the result size and lower data traffic. These could help the server to guess the user's sub-type of interest with a high degree of confidence and lead to a loss of privacy. To tackle this problem, our approach equalizes the amount of data in each row of the sub-type in a specific class by adding "null" to all the sub-types which are not equal to the maximum sub-type size in that class. The main advantage of our "null" solution is that if the user is looking for a sub-type with less data, the PIR computation overhead on the

client-side is reduced (when receiving the first null in the decoding process, the computation process stops), and if the user is looking for a sub-type with more data, the PIR computation cost on the client-side increases, without losing privacy.

Finally, in our approach the amount of data is different in different classes (see Figure 1), unlike [9]. As a result, our PIR computation cost depends on the specific class that the user requests. If the user searches for a class with more data, the PIR computation cost increases. If the user searches for a class with less data, the PIR computation cost is reduced.

By considering these three improvements, if the user sends a query to the database for the sub-type of data in each class, the response which is returned to the user not only needs less computation time for decoding, but also does not need to be filtered on the client side to remove non-requested data.

Class	Category Types					
Class-0 00	Sub-type 0 0000	data1	data2			
	Sub-type 1 0001	data 1	-			
	Sub-type 2 0010	data1	data2			
	Sub-type 3 0011	-	-			
	Sub-type 4 0100	-	-			
	Sub-type 5 0101	data1	data2			
	Sub-type 6 0110	data1	data2			
	Sub-type 7 0111	-	-			
	Sub-type 8 1000	-	-			
	Sub-type 9 1001	data1	-			
Class-1 01	Sub-type 0 0000	-				
	Sub-type 1 0001	-				
	Sub-type 2 0010	-				
	Sub-type 3 0011	-				
	Sub-type 4 0100	-				
	Sub-type 5 0101	-				
	Sub-type 6 0110	-				
	Sub-type 7 0111	-				
	Sub-type 8 1000	-				
	Sub-type 9 1001	-				
Class-2 10	Sub-type 0 0000	-	-	-	-	-
	Sub-type 1 0001	data1	data2	data3	data4	data5
	Sub-type 2 0010	-	-	-	-	-
	Sub-type 3 0011	data1	data2	-	-	-
	Sub-type 4 0100	data1	-	-	-	-
	Sub-type 5 0101	data1	data2	data3	data4	data5
	Sub-type 6 0110	data1	-	-	-	-
	Sub-type 7 0111	data1	data2	data3	data4	-
	Sub-type 8 1000	-	-	-	-	-
	Sub-type 9 1001	-	-	-	-	-
…	…					

Figure 1 Sample of the relationship between data, sub-type and classes, as saved in the database.

3.2. The Proposed PIR Scheme

Our PIR protocol has two phases. The first phase is the pre-processing phase in which the whole protocol becomes ready to use, on the server side and also on the client side. In the future, if the client decides to change the level of her privacy, or any changes occur on the server side, this phase can be repeated. The second phase is the execution phase in which the user sends her request to the server. Her request contains the class of data which she searches concatenated with the sub-type category.

Pre-processing Phase contains the following steps:

1. Given a chosen level of the user's privacy, "Class", "sub-type" and "data" category are applied to the database.
2. The "class" and "sub-types" are defined to have a number based on their specific categories. As shown in Figure 1, for example, Class-1 is considered 01 and Sub-type-3 is considered 0011. Note that in this Figure we just showed "10" different "sub-types" for each "class". This depends on the number of different sub-types of data in the database and also on the level of the user's privacy which is applied in step 1.
3. Each database index will be the "class" concatenated with the "sub-type". For example, in Figure 1, the database index for the Class-1||sub-type-3 is considered as 010011.

Execution Phase contains the following steps:

1. The user chooses the sub-type of her interest from the list suggested by the application based on her privacy level. For example, she is looking for Class-1||sub-type-3.
2. The proposed application provides the user's request, which is an index of the database, and sends it to the server in a way that is hidden from the server. In this example, the request is 010011 which refers to Class-1||sub-type-3.
3. The specific row of database is retrieved from the database and the data present in this row are transmitted back to the user.
4. The user decodes the results and the results are shown on her smartphone.

4. The Proposed Privacy-Preserving Protocol for LBS

The main goal of cryptographic protocols in nearby places is to be able to detect nearby places automatically while the user's location privacy is considered in the location-based service (LBS) application. Our proposed protocol uses private information retrieval (PIR) to achieve this purpose.

4.1. Problem Statement

Alice has her location as her secret. Alice wants to use a LBS application to search and find nearby places of interest. We propose a protocol that allows Alice to find nearby places for which she is looking. However, the LBS that helps Alice to find her nearby place does not learn Alice's location. Alice can send a request to the LBS's database to fetch her nearby places of interest without the LBS being aware of what Alice fetched by using private information retrieval (PIR). Most of the previous PIR schemes are not acceptable in LBS applications because of their use of secure hardware. The focus of this section is to solve the PIR-based LBS issues by offering a practical PIR without using secure hardware or a trusted third party and lower the computational cost on the client side in the smartphone's application. At the end of this protocol, the proposed application should list the POIs that meet Alice's search criteria or show her that there is no POI in the selected area.

4.2. The Proposed Privacy-Preserving Protocol for LBS

We first informally describe our proposed protocol via an example. Suppose Alice is located in Ottawa and she wants to look for a specific type of POI, for example a restaurant, near Bank Street. Since she is privacy conscious, she sets her cloaking area to be a 10 km MGRS grid square (see section 4.3). The client

application sends the requested cloaking area to the server. At the same time, the PIR allows the client application to identify which part of the cloaking area has the restaurant, without the server being informed which part is retrieved. All entries in the POI database are indexed by their MGRS block concatenated with the POI type. The row that contains the restaurant(s) is retrieved from the selected MGRS grid square on the database and the results are sent back to Alice. The client application decodes the results and sorts the results, and the nearest restaurants are shown on her phone's local map.

Our protocol follows [9] hybrid solution that uses PIR to preserve the privacy of the user's query and a cloaking scheme in order to make the PIR scheme practical and reduce the computational cost of PIR. The benefits of the hybrid solution are as follows: the location of the user remains secret from the LBSP to a reasonable privacy level chosen by the user without depending on the other users in the selected area; to calculate the cloaking area or cryptographic algorithms we do not need to have a trusted third party; and the computational overhead of the PIR scheme is practical.

Our proposed protocol has two improvements compared to [9]. First, due to the user's request for a specific POI, our proposal categorizes the cloaking area in the database into the POI types. Thus, when the user asks for her POI in her selected cloaking area, the results that are returned to her are of the type that she is looking for. Therefore, our protocol on the client is not required to filter the block of different types of POIs to identify the POI that the user requested. This reduces the computational costs, and saves the battery and data usage on the smartphone. For example, if there are no restaurants near the user, she does not need to wait to decode all POIs in that cloaking area and then filter the restaurant to find that actually the answer is "null".

Second, we propose a new technique based on a static grid-based approach for defining our cloaking area and mapping our POIs to a cloaking area, unlike the approach proposed in [9] which uses the Various-size-grid Hilbert Curve (VHC) technique [35] (see section 4.4).

Our proposed protocol describes how the POI database is initialized and how the protocol generates a cloaking area around the user's exact location, and executes a PIR query on the contents of the requested cloaking area. We name our phases similarly to [9] to highlight the similarities and differences between our phases. Note that each POI consists of 300 bytes that includes longitude and latitude coordinates, name, exact address, the phone number, website address, etc.

The pre-processing phase contains the following steps:

1. An appropriate static grid system is applied on the geolocation plane.
2. POIs are categorized based on their type and saved in the LBS's database.
3. A row of database refers to a cloaking area concatenated with the POI type.

The execution phase contains the following steps:

1. The user selects the area of her interest; it could be her current location as determined through GPS, or some other location that the user may be traveling to in the future.
2. The user selects a preferred level of privacy.
3. The user's corresponding cloaking area is calculated based on the level of her chosen privacy.
4. The user chooses the POI type(s) from the suggested list provided by the client application.
5. The client application sends the cloaking area to the server. Also, the client application identifies which portion of the cloaking area contains the POI type(s), in a way that is hidden from the server.
6. The server receives the request, and finds the database portion corresponding to the cloaking area. A block of rows is retrieved from this portion based on the user's specified POI type. The POIs present in these rows are transmitted back to the client application.
7. The client application decodes and sorts the results, and the nearest POIs are shown on her phone's local map.

4.3. Grid-based Cloaking

In our proposed protocol, the client application extracts the user's location via cell towers, Wi-Fi, or GPS, and it calculates the user's cloaking area by using the military grid reference system (MGRS) technique. MGRS is a geo-coordinate standard for locating points on the Earth [36]. The Earth is divided into grid squares with sizes of 0.1 km, 1 km, 10 km, 100 km, etc., based on the level of accuracy and degree of precision. Our proposed protocol uses MGRS to help ensure the user's location privacy. Each MGRS block is considered as a block in the database that is categorized based on POI type.

considered to be nearby if they are located in the same MGRS block as the user. The user's location privacy level increases if she chooses a larger MGRS block. However, a larger MGRS block includes more POIs, and it affects the computational cost of our proposed protocol. We will discuss this issue in more detail in the following section. Figure 2 shows the different levels of MGRS blocks for the Ottawa, Ontario, area [37].

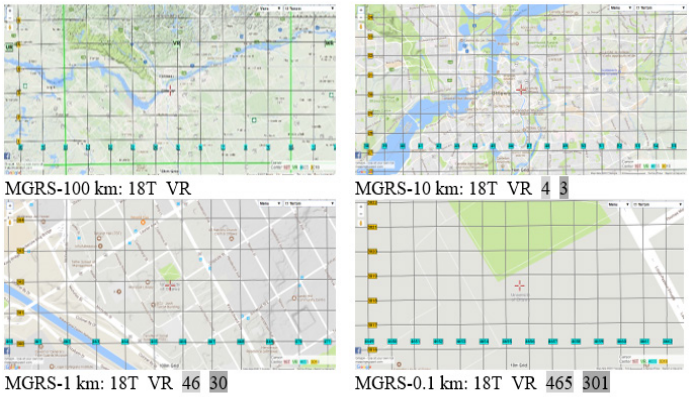
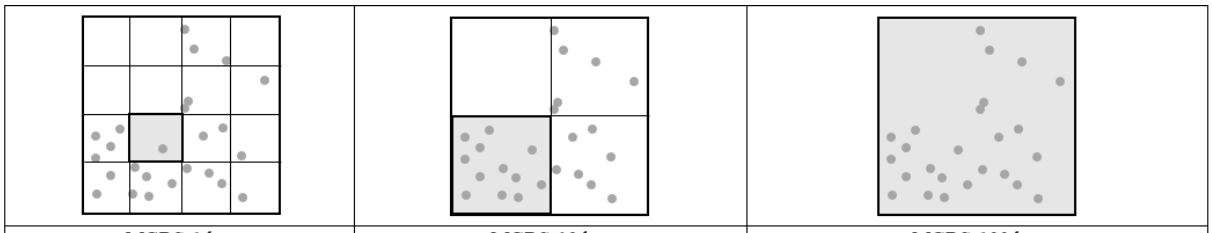


Figure 2 Different MGRS Levels [37]



MGRS-1 km

MGRS-0.1 km	POI type	POIs
18TVR 46 30	0000	null
	0001	null
	0010	POI1
	0011	null
	0100	null
	0101	null
	0110	null
	0111	null
	1000	null
	1001	null

MGRS-10 km

MGRS-1 km	POI type	POIs		
18TVR 4 3	0000	null	null	null
	0001	null	null	null
	0010	POI1	POI2	null
	0011	POI1	null	**null**
	0100	POI1	null	null
	0101	POI1	POI2	POI3
	0110	POI1	null	null
	0111	POI1	null	null
	1000	null	null	null
	1001	POI1	POI2	POI3

MGRS-100 km

MGRS-10 km	POI type	POIs			
18TVR	0000	POI1	POI2	null	null
	0001	POI1	POI2	POI3	null
	0010	POI1	POI2	null	null
	0011	POI1	null	null	null
	0100	POI1	POI2	POI3	null
	0101	POI1	POI2	POI3	null
	0110	POI1	null	null	null
	0111	POI1	POI2	POI3	POI4
	1000	POI1	POI2	null	null
	1001	POI1	POI2	POI3	POI4

Figure 3 Illustration of the relationship between the MGRS block, POI types, and POIs as saved in the LBS database.

4.4. Location Cloaking

The first step in the location cloaking phase is to apply different MGRS levels on the geographic region, such as Canada and the U.S.. Then, the user's cloaking area is calculated based on the user's selected MGRS level and her current location or the location of interest. POIs are considered to be nearby if they are located in the same cloaking area.

The selected MGRS level must be large enough to achieve the privacy of the user's location within the requested cloaking area, but simultaneously it must be small enough to reduce the computational overhead on the smartphone application to process the results, and also to reduce the communication overhead to transfer the result via the wireless data traffic.

In order to map POIs to a cloaking area, [9] used the Various-size-grid Hilbert Curve (VHC) technique. [9] chose VHC because it could solve the issue of density of POIs based on the geographic region. For example, the data traffic cost increases if the selected area has a high density of POIs (within a city). On the other hand, if the selected area has a lower density of POIs (within a countryside), then the result size decreases and the server is able to estimate the location of the user, which leads to a loss of privacy. VCH can solve this problem by creating a different-sized cloaking area based on the density of POIs. However, this solution has the disadvantage of receiving a list of POIs which may or may not be useful for the user. If the selected area has no POI that the user is looking for, she still has to wait for the client application to calculate the result, which is based on all POIs in a selected VCH region, and then show the result which is actually "null". This can cause a high computational cost on the client side application.

To manage the computational cost on the smartphone application based on the density of the POIs, and prevent high computational cost in the lower density POIs area, we propose a new technique to map POIs to a cloaking area based on the MGRS fix-sized blocks. First of all, we categorize POIs based on their types in each MGRS block. In Figure 3 we consider ten POIs types per MGRS block (see Appendix A) [7]. Then, each row of the database refers to the MGRS block concatenated with the POI type. As [9] mentioned, the density of POIs varies by geographic area. Therefore, each row of the defined database has a variable size.

To protect the privacy of the user's location and prevent the server from guessing which POI type is fetched by the user, we need to equalize the number of POIs in each selected cloaking area. Therefore, if the number of POIs in one POI is not the same as the maximum POI size in the selected cloaking area, the rest of the row must be set to "null". By this technique, our PIR client side computational cost relies on the location of the user and the level of selected MGRS. If the user's location has low POI density, the PIR client side computation time will decrease. If the user's location has high POI density, the PIR computational cost on the client side will increase. Note that the server cannot observe the differences between computational costs for queries in different locations, because we equalized the number of POIs in the selected cloaking area. Otherwise, the server which is able to observe the difference between computational cost based on different user's queries, could guess the user's location. Figure 3 shows an example of the POI density in the selected area based on different levels of MGRS and illustrates the relationships among the MGRS block, POI types, and POIs as saved in the LBS database.

5. Security Analysis of the Proposed Protocol

The user can use her current location or a location that she wishes to visit in the future. This feature adds one more level of privacy in our protocol because the observer or the location-based service provider (LBSP) is not aware of whether the requested MGRS block corresponds to the user's current location. Therefore, our proposed protocol has two kinds of privacy: first, protection of the user's location privacy within the requested MGRS block, and second, the LBSP or an observer does not know whether the

request is the user's current location at the request's time. In both cases, our main goal is to protect the user's location privacy against the LBSP and any other observer.

5.1. Threat model

We can apply our proposed PIR protocol to all existing block-based PIR schemes (CPIR and IT-PIR). In this section, we use the IT-PIR (multi-server) to describe our threat model (as was done by [9]). The primary assumption in IT-PIR schemes is that servers must not communicate with each other to breach the privacy of users' queries. Under this assumption, the IT-PIR protocol itself has been proven secure in [10, 15, 27]. Given the cryptographic security of the IT-PIR scheme, we review the security of our PIR protocol, as well as its security against passive and active adversaries in the following sections.

5.2. Security Analysis

Claim 5.2.1. If B is an MGRS block of level L, chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against a *passive adversary* in the block B.

Passive adversary. An external observer or a malicious LBSP who has access to the data exchanged between the user and the database along the communication channel but cannot change the data.

Justification. Our proposed protocol calculates the portion of the database in which Alice wants to find a nearby place of interest. It depends on the level of privacy she selected, such as MGRS-100 m. Therefore, Alice's location privacy and query are limited to the requested portion of the database. The number of POIs in each type of an MGRS block is set to be equal to the maximum number of POIs in that MGRS block by adding "null" to the ends of the other types in our proposed database, Thus, the passive adversary cannot guess which type of POI was fetched by Alice. In other words, if Alice's type of POI changes while she is still in the same MGRS block, then for a new request, Alice will send the same query for the same MGRS block of the database.

Because Alice selected her level of privacy to be, for example, a block of MGRS-100 m, it is impossible for the passive adversary to detect her movement as long as she is in the same MGRS block. Therefore, a correlation attack is actually unachievable since Alice will always send the same query for a given privacy level. Additionally, if Alice knows that she is going to move, she should choose a larger MGRS block that includes both her current location and her next location. Thus, her movement will not be detectable.

The PIR scheme provides the security for our proposed privacy-preserving protocol for LBS against a passive adversary. The only information which the passive adversary can access is the identity of the user. If the user's identity protection is required, we can use TOR. Additionally, the user's query contents and the response of the database can be protected against a malicious observer by applying end-to-end encryption techniques, for example, TLS (transport layer security) through the wireless communication channel. Both schemes (TOR and TLS) are optional for the user, as they slow down the protocol and cause an additional computational cost.

Claim 5.2.2. If B is an MGRS block of level L chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against an *active adversary* in the block B.

Active adversary. A malicious external observer who has access to data exchanged between the user and the database along the communication channel and can delete, insert or modify data. The LBSP is considered to be trusted in Claim 5.2.2.

Justification. Message reordering attack. During this attack, the active adversary attempts to delay and/or reorder requests or responses to mix up results and the communication. Note that in our proposed protocol, if the results of multiple queries are not received in proper sequence, it has no effect on the server or Alice since each result holds the requested information. The adversary also gains no information about Alice's query or location.

Justification. Message tampering attack. Alice will not be able to verify false responses if the adversary starts to send them. Therefore, a DoS (denial-of-service) attack is possible. However, in using this attack, the active adversary will not learn any more information about Alice's query or location, which is the main focus here. Note that this attack can be prevented by using TLS over the communication channel.

Justification. Message insertion or deletion attack. If an active adversary tries to delete or insert data from the server's response or Alice's request, it can cause a DoS attack. The adversary does not receive any information about Alice's request or location. Again, TLS can be used on the communication channel to prevent this attack, if desired.

Justification. Message replay attack. In this attack, if an active adversary starts a replay attack against Alice or the server, it does not affect either of them. The server responds to the requests, and Alice can easily drop multiple responses with the same information. The adversary will not receive any information about Alice's query or location. As with the previous attacks, we can use TLS to prevent this attack as well.

Claim 5.2.3. If B is an MGRS block of level L chosen by Alice, and T is the type of POI that Alice is searching for, our proposed privacy-preserving protocol for LBS is secure against a *malicious server* in the block B.

Malicious Server. A malicious server refers to an LBSP that attempts to insert new messages, modify, or delete messages in response to the user.

Justification. If the malicious server sends a false response, does not return a response, or sends additional messages with the

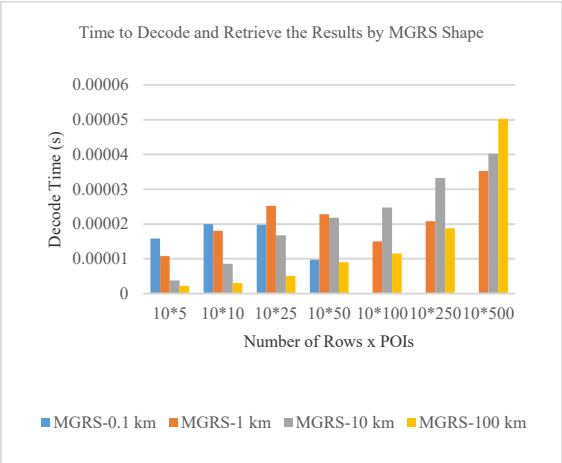


Figure 4 Comparison of time to decode and retrieve the results, by MGRS shape at the client side. It shows the computation time for queries on one MGRS block (ten POI types) for different number of POIs (each POI consists of 300 bytes).

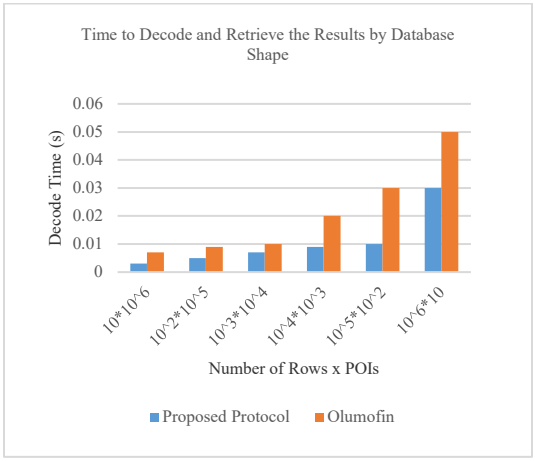


Figure 5 Comparison of decode and result retrieval time by database shape at client side in our proposed protocol and in Olumofin. The results show the computation cost for different-shaped database for queries on a 3 GB database (each POI consists of 300 bytes).

response, it can cause a DoS attack against the user. However, this attack does not enable the malicious server to learn any information about the location of the user or her query. The only thing that could help the malicious server find information about the location of the user is the content of the query, which is protected by using the PIR scheme within the MGRS block.

6. Experimental evaluation and results

We implemented a prototype of our proposed protocol based on an open source PIR protocol called Percy++ [15, 16, 38, 39, 40, 41]. We ran our C++ prototype on a virtual machine with Ubuntu Linux operating system and an Intel® Core™ i7-8550U CPU @ 1.80 GHz, 4GB RAM. We followed all assumptions of [9] in our implementation to compare our results with their approach. We randomly generated and distributed ten million POIs within Canada and the U.S. [7]. Each POI consisted of 300 bytes that included the longitude and latitude coordinates, name, exact address, phone number, website address, etc., of the POI [9]. We set the number of the databases to two to use the Percy++ PIR [9]. For each of these, we applied MGRS to our generated map to create four databases for four levels of user privacy: MGRS-0.1 km, MGRS-1 km, MGRS-10 km, MGRS-100 km.

Recall that the main goal of this section is to decrease the decode time on a smartphone to make the PIR scheme practical for resource-constrained hardware. The decoding time has a direct correlation with the number of POIs in each MGRS block. As we defined the fixed number of POI types for different levels of MGRS blocks (ten types), the PIR decoding time on the smartphone depends on the number of POIs in each type. Therefore, if the number of POIs in a type increases, the decoding time increases. We equalized the number of POIs in all types of MGRS blocks by adding "null" entries, as explained above. Therefore, there can be types that have no data or less than the maximum POI type. This improves the decoding time compared to a type that has the maximum number of POI. As observed in Figure 4 , when the user's level of privacy is increased (larger MGRS block), the probability of obtaining more POIs per type increases. Therefore, an MGRS block with a maximum of five POIs per type has faster data retrieval compared to 500 POIs per type. However, there are some exceptions. For example, MGRS-0.1 m required less computational time with 50 POIs per type than with five POIs per type. These exceptions can happen if the type that was requested has fewer POIs than 50, and the rest of the data in that type is "null" (to make the total number of POIs equal to the maximum number of POIs in the requested MGRS block). During the decode time, whenever the first "null" block is detected, the decode operation stops and results are returned to the user.

We ran our prototype using different levels of MGRS blocks, and we generated 100 random requests to calculate the average time to decode and retrieve the results. As observed in Figure 4, the probability that we had 500 POIs from one type in an MGRS-0.1 km was equal to zero. This means, for example, that in an area of 0.1 km^2, there cannot be 500 banks. The opposite scenario may also occur. The probability that we have five POIs of one type in an MGRS-100 km block is rare. Considering this, the MGRS-10 km is the best choice if we want to show the results for an MGRS block with various numbers of POIs/types.

In our proposed protocol, the number of rows in each MGRS block is the same as the number of POI types, which was set to ten types per MGRS block, regardless of the size of the MGRS block. Thus, if our MGRS block becomes larger, the number of types (rows) does not change, but the number of POIs per type increases. We also have different numbers of POIs per type. Thus, in an MGRS block, we have some types that have no data or less than the maximum number of POIs in that MGRS block. This is important because we want to compare our results with [9], and due to the different definition of privacy and the reasons we just mentioned, it is difficult to give an exact comparison.

However, to show the performance of our proposed protocol compared to [9], we followed their implementation by setting the privacy level equal to one. Thus, we considered one large MGRS block that covered Canada and the U.S., and the number of POI types was set equal to the number of rows in [9]. Our implementation results are based on a database of ten million POIs. Figure 5 shows the time for decoding and retrieving the results for various numbers of rows and POIs.

As observed in Figure 5, our performance is approximately 50% better than that of [9] because our method considers the POI type and uses the MGRS which applies a fix-sized cloaking area and a variable-sized block to the database. In our proposed protocol, the user receives exactly the type of POI that she was looking for. By increasing the number of POIs per row, the decode and retrieve time increases in [9] protocol because after decoding the rows of the database, the results must be filtered to show the POI that the user was looking for.

As stated in section 4.3, due to the four levels of MGRS (0.1 km, 1 km, 10 km and 100 km), our database required four different configurations. This was the only disadvantage of our proposed protocol compared to [9]. This could increase processing on the server when adding or removing POIs from the databases (for example, when a restaurant closes or a new one opens in a specific MGRS block).

7. Limitations of our Proposed Protocol

In general, there may be a case in which the user will not find a reasonable POI in the requested cloaking area. Therefore, she may wish to search further in a larger MGRS block (i.e., in a broader geographical area). When this happens, the user's privacy does not decrease in our proposed protocol; it is still guaranteed to the level of the original cloaking area.

As mentioned in section 4.3.3, due to the four levels of MGRS (100 km, 10 km, 1 km and 0.1 km), we required four different configurations for our database. This is a disadvantage of our proposed protocol compared to [9]. This could increase processing on the server when adding or removing POIs from the databases (for example, when a restaurant closes or a new one opens in a specific MGRS block).

Modern smartphones with multi-core processors may be able to handle the 1.5 GB database for Canada and the U.S. that is used in [9] evaluation section, as well as the most recent 3 GB location database [7] that we used in our implementation. However, we should mention that not all people have the most recent smartphones and so our proposal, which reduces computational cost on the client by almost 50%, may be of particular interest for such environments.

8. Conclusion and Future Work

In this paper, we presented a privacy-preserving protocol to help the user search for nearby places of interest while protecting her location's privacy by using PIR. For this purpose, we first proposed a block-based PIR scheme to decrease the computational overhead on smartphone applications [8]. We demonstrated that by applying our PIR scheme to the LBS, the computational overhead on the client side was reduced by approximately 50% compared to that reported in a previous work [9]. This reduction is valuable for the implementation of PIR in smartphone applications with limited resources. We demonstrated that our proposed LBS protocol is secure against active and passive attacks, as well as against a malicious server that tries to identify information about the user's query and location.

Our approach of retrieving the specific POIs in a cloaking area consumes less computational cost compared with the naive approach of asking the user to download the entire contents of the cloaking area and extract POIs locally. Simultaneously, the user's location privacy is not compromised since the user requests the same cloaking area as if she was requesting the entire contents of the cloaking area. This is a great benefit that reduces the cost of the wireless communication and also the memory usage on the smartphone.

There exist a number of interesting directions for our LBS privacy future work. First, our implementation results are based on [15]; in order to improve the computation, cost of our proposed protocol, we could develop it based on the higher performance block-based PIR such as the hybrid PIR which is proposed in [13]. Second, our proposed protocol could be extended by supporting more complex types of queries. Third, our proposed protocol could be combined with the Vehicle-to-Infrastructure (V2I) and the Vehicle-to-Vehicle (V2V) communication to help the user find her nearby places while she is driving her car. In this scenario, the user receives the latest update about the nearby places for her response from other vehicles or street infrastructure such as traffic lights and signs instead of a solid database. To update information about places and their availability constantly, we could use a Blockchain infrastructure in which other vehicles or street infrastructure are able to update their recent observations about the places. For example, all the infrastructure components on the street are connected to the Blockchain and every change that occurs appears in the block. Now if the user is looking for nearby parking, the traffic light could let her know the nearest parking lot and if there is any spot available by checking the updated list on the Blockchain.

References

[1] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: Anonymizers are not necessary" In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 121–132, 2008. https://doi.org/10.1145/1376616.1376631

[2] W. Hao, and Y-C. Hu, "Location Privacy with Randomness Consistency" In Proceedings on Privacy Enhancing Technologies, 62–82, 2016. https://doi.org/10.1515/popets-2016-0029

[3] U. Hengartner, "Hiding location information from location-based services" In International Conference onMobile Data Management, 268–272, 2007. https://doi.org/10.1109/MDM.2007.56

[4] M. Hezaveh, and C. Adams, "Privacy Preserving Discovery of Nearby-Friends" In E-Technologies: Embracing the Internet of Things, Lecture

Notes in Business Information Processing, (289), 41–55, 2017. https://doi.org/10.1007/978-3-319-59041-7_3

[5] A. Khoshgozaran, C. Shahabi, and H. Shirani-Mehr, "Location privacy: going beyond K-anonymity, cloaking and anonymizers" Knowledge and Information Systems, 26(3), 435–465, 2011. https://doi.org/10.1007/s10115-010-0286-z

[6] Y, G. Kim, J. Kong, and S. W. Chung, "A Survey on Recent OS-level Energy Management Techniques for Mobile Processing Units" IEEE Transactions on Parallel and Distributed Systems, 2388–2401, 2018. https://doi.org/10.1109/TPDS.2018.2822683

[7] Factual, Global Places-Schema, https://my.factual.com/data/t/places, last accessed 2018/11/05.

[8] M. Hezaveh and C. Adams, "A PIR scheme to improve the computation cost on the client-side of smartphone application" In IEEE 31th Canadian Conference on Electrical and Computer Engineering, 1–4, 2018. https://doi.org/10.1109/CCECE.2018.8447708

[9] F. Olumofin, P. K. Tysowski, I. Goldberg, U. Hangartner, "Achieving efficient query privacy for location-based services" International Symposium on Privacy Enhancing Technologies Symposium, 93–110 2010. https://doi.org/10.1007/978-3-642-14527-8_6

[10] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval" In Proceedings of the 36th Annual Symposium on the Foundations of Computer Science, 41–50, 1995. . https://doi.org/10.1145/293347.293350

[11] D. Lin, E. Bertino, R. Cheng, and S. Prabhakar, "Position transformation: a location privacy protection method for moving objects" In Proceedings of the SIGSPATIAL ACM GIS 2008 Inter- national Workshop on Security and Privacy in GIS and LBS, 62–71, 2008. https://doi.org/10.1145/1503402.1503414

[12] D. Riboni, L. Pareschi, and C. Bettini, "Privacy in georeferenced context-aware services: A survey" Privacy in Location-Based Applications, 151–172, 2009. https://doi.org/10.1007/978-3-642-03511-1_7

[13] C. Devet and I. Goldberg, "The best of both worlds: Combining information-theoretic and computational PIR for communication efficiency" In International Symposium on Privacy Enhancing Technologies Symposium, 63–82, 2014. https://doi.org/10.1007/978-3-319-08506-7_4

[14] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: the second-generation onion router" In Proceedings of the 13th conference on USENIX Security Symposium, 2004. https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf

[15] I. Goldberg, "Improving the robustness of private information retrieval" In Proceedings of the IEEE Symposium on Security and Privacy, 131–148, 2007. https://doi.org/10.1109/SP.2007.23

[16] C. Aguilar-Melchor and P. Gaborit, "A lattice-based computationally-efficient private information retrieval protocol" In Western European Workshop on Research in Cryptology, 2007. https://eprint.iacr.org/2007/446.pdf

[17] R. Sion, and B. Carbunar, "On the computational practicality of private information retrieval" In Proceedings of the Network and Distributed Systems Security Symposium, 2007. https://zxr.io/research/sion2007pir.pdf

[18] W. Gasarch, "A survey on private information retrieval" The Bulletin of the EATCS, 82, 72–107, 2004. https://www.researchgate.net/profile/William_Gasarch/publication/266280304_A_survey_on_private_information_retrieval/links/5705098d08ae74a08e270e57.pdf

[19] B. Chor and N. Gilboa, "Computationally private information retrieval" In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing, 304–313, 1997. https://doi.org/ 10.1145/293347.293350

[20] B. Chor, N. Gilboa, M. Naor, "Private Information Retrieval by Keywords" Technion-IIT, Department of Computer Science, 1997.

[21] E. Kushilevitz, and R. Ostrovsky, "Replication is not needed: single database, computationally-private information retrieval" In Proceedings of the 38th Annual Symposium on Foundations of Computer Science, 364–373, 1997. https://doi.org/10.1109/SFCS.1997.646125

[22] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylog communication" In International Conference on the Theory and Applications of Cryptographic Techniques, 402–414, 1999. https://doi.org/10.1145/258533.258609

[23] C. Gentry, and Z. Ramzan, "Single-database private information retrieval with constant communication rate" International Colloquium on Automata, Languages, and Programming, 803–815, 2005. https://doi.org/10.1007/11523468_65

[24] E. Kushilevitz, and R. Ostrovsky, "One-way trapdoor permutations are sufficient for non-trivial singleserver private information retrieval" In International Conference on the Theory and Applications of Cryptographic Techniques, , 104–121, 2000. https://doi.org/10.1007/3-540-45539-6_9

[25] H. Lipmaa, "An Oblivious Transfer Protocol with Log-Squared Communication" In International Conference on Information Security, 314–328, 2005. https://doi.org/10.1007/11556992_23

[26] C. Aguilar-Melchor, J. Barrier, L. Fousse, and M.O. Killijian, "XPIR: Private information retrieval for everyone" In Proceedings on Privacy Enhancing Technologies, 155–174, 2016. https://doi.org/10.1515/popets-2016-0010

[27] A. Beimel and Y. Stahl, "Robust information-theoretic private information retrieval" Journal of Cryptology, 20(3), 295–321, 2007. https://doi.org/10.1007/s00145-007-0424-2

[28] Y. Gertner, S. Goldwasser, T. Malkin, "A Random Server Model for Private Information Retrieval" In 2nd International Workshop on Randomization and Approximation Techniques in Computer Science, 200–217, 1998. https://doi.org/10.1007/3-540-49543-6_17

[29] S. Wang, X. Ding, R. H. Deng, and F. Bao, F, "Private information retrieval using trusted hardware" European Symposium on Research in Computer Security, 49–64, 2006. https://doi.org/10.1007/11863908_4

[30] E. Fung, G. Kellaris, and D. Papadias, "Combining Differential Privacy and PIR for Efficient Strong Location Privacy" International Symposium on Spatial and Temporal Databases, 295–312, 2015. https://doi.org/10.1007/978-3-319-22363-6_16

[31] S. Papadopoulos, S. Bakiras, and D. Papadias, "Nearest neighbor search with strong location privacy" In Proceedings of the VLDB Endowment, 3(1-2), 619–629, 2010. https://doi.org/10.14778/1920841.1920920

[32] G. Ghinita, "Understanding the privacy-efficiency trade-off in location based queries" In Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, 1–5, 2008. https://doi.org/10.1145/1503402.1503404

[33] P. Paillier, "Public key cryptosystems based on composite degree residue classes" In International Conference on the Theory and Applications of Cryptographic Techniques, 223–238, 1999. https://doi.org/10.1007/3-540-48910-X_16

[34] A. Shamir, "How to Share a Secret" Communications of the ACM, 22(11), 612–613 1979. https://doi.org/10.1145/359168.359176

[35] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, "CAP: A Context-Aware Privacy Protection System For Location-Based Services" In 29th IEEE International Conference on Distributed Computin Systems, 49–57, 2009. https://doi.org/10.1109/ICDCS.2009.62

[36] S. Hemisphere, "Northern Hemisphere" Ann Arbor 1001, 2006. (see also: https://en.wikipedia.org/wiki/Military_Grid_Reference_System)

[37] Mapping support, https://mappingsupport.com/p/gmap4.php?tilt=off&mgrs=14SPG34308382&z=5&t=t1, last accessed 2018/11/05.

[38] C. Devet, I. Goldberg, and N. Heninger, "Optimally Robust Private Information Retrieval" In USENIX Security Symposium, 269–283, 2012. https://eprint.iacr.org/2012/083.pdf

[39] I. Goldberg. Percy++ project on SourceForge, http://percy.sourceforge.net/. last accessed 2018/11/05.

[40] R. Henry, F. Olumofin, I. Goldberg, "Practical PIR for Electronic Commerce" In Proceedings of the 18th ACM conference on Computer and communications security, 677–690, 2011. https://doi.org/10.1145/2046707.2046784

[41] W. Lueks, I. Goldberg, "Sublinear Scaling for Multi-Client Private Information Retrieval" In International Conference on Financial Cryptography and Data Security, 168–186, 2015. https://doi.org/10.1007/978-3-662-47854-7_10

Appendix A

A global, open, collaborative, standardized points of interest database provided by Factual [7].

	POI Type	Canada	US
1	Automotive	41913	617664
	Automotive, maintenance and repair	27267	285077
	Automotive, maintenance and repair, tires	0	73587
	Automotive, car parts and accessories	6917	79292
	Automotive, car dealers and leasing, car dealers	7729	105690
	Automotive, car dealers and leasing, used car	0	74018
2	Businesses and services, financial	35914	348814
	Businesses and services, financial, bank and finance, bank and credit union	14886	128993
	Businesses and services, financial, financial planning and investments	7919	91282
	Businesses and services, financial, access and bookkeeping	13109	128539
3	Businesses and services	93421	1181684
	Businesses and services, personal care, beauty salons and barbers	26148	325314
	Businesses and services, shipping freight, and material transportation	6232	51966
	Businesses and services, insurance	13687	224441
	Businesses and services, legal, attorney and law offices	12438	226907
	Businesses and services, real estate, real estate agents	9869	140201
	Businesses and services, Telecommunication services	6667	47031
	Businesses and services, computers	9684	92043
	Businesses and services, printing, copying and signage	8696	73781
4	Businesses and services, home improvement	138297	1403076
	Businesses and services, home improvement	93762	962116
	Businesses and services, home improvement, contractors	25612	260638
	Businesses and services, home improvement, ventilating and air conditioning, heating	5970	74885
	Businesses and services, home improvement, plumbing	6279	58550
	Businesses and services, home improvement, electrician	6674	46887
5	Community and government	57185	839554
	Community and government, organization and associations	14688	136389
	Community and government, education and secondary schools Primary and secondary school	14907	192245
	Community and government, day care and preschools	6859	80547
	Community and government, religious, churches	14531	362889
	Community and government, public and social services	6200	67484
6	Healthcare	51359	969517
	Healthcare, dentists	18683	230012
	Healthcare, pharmacies	11840	62652
	Healthcare, hospitals, clinics and medical centers	8221	155416
	Healthcare, physicians	12615	521437
7	Retail	81534	758818
	Retail, furniture and décor	8656	96082
	Retail, fashion, shoes	5946	61335
	Retail, fashion, clothing and accessories	22886	193354
	Retail, fashion, jewelry and watches	6115	67001
	Retail, construction supplies	5776	42307
	Retail, supermarkets and groceries	10862	98231
	Retail, food and beverage, beer, wine and spirits	6014	54942
	Retail, convenience stores	9610	100149
	Retail, glasses	5669	45417
8	Social	179478	1911585
	Social, food and dining, restaurants	101714	1017499
	Social, food and dinning, restaurants, fast food	17207	236356
	Social, food and dinning, restaurants, dining	16033	0
	Social, food and dinning, cafes, coffee and tea houses	15083	117367
	Social, food and dinning, restaurants, pizza	12888	136018
	Social, food and dinning, restaurants, American	0	225245
	Social, food and dinning, restaurants, Chinese	5724	0
	Social, Bars	10829	179100
9	Transportation	17906	206107
	Transportation, gas stations	11783	161466
	Transportation, taxi and car services, car and truck rentals	6123	44641
10	Travel	27937	177682
	Travel, travel agents and tour operators	6505	35613
	Travel, lodging, hotel and motels	21432	142069

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 3, 38-46 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Stress Level Classification Using Heart Rate Variability

Tlija Amira[1,2,*], Istrate Dan[1], Badii Atta [3], Gattoufi Said[2,4], Bennani Az-eddine[5,6], Wegrzyn-Wolska Katarzyna[7]

[1]*BMBI UMR 7338, UTC, 60200, Compiegne, France*

[2]*SMART Lab, ISG TUNIS, 2000, Bardo, Tunisia*

[3]*Chair of Secure Pervasive Technologies, Computer Science Department, University of Reading, RG6 6AH, United Kingdom*

[4]*CEMIS-University of Nizwa-Sultanate of Oman, Oman*

[5]*TSH COSTECH, UTC, 60200, Compiegne, France*

[6]*Neoma Business School, 5100, Reims, France*

[7]*Allianstic, EFREI PARIS, 94800, Villejuif, France*

ARTICLE INFO

Article history:
Received: 20 December 2018
Accepted: 08 April, 2019
Online: 09 May, 2019

Keywords:
Heart Rate Variability (HRV),
RR interval,
Wearables sensors,
IoT,
Support vectors machines,
Classification

ABSTRACT

The research programme reported in this paper is set within the framework of our research under the theme of ICT support for Active Healthy Ageing (AHA). This longitudinal empirical research is focused on the study of the impact on the management of cardio-vascular disease if supported by sustained health monitoring using wearable connected devices. One of the key objectives is stress monitoring and its classification during the daily routine of life thus enabling psycho-physiological monitoring to study the correlation between emotional states including variable stress levels and the evolution and prognosis of cardiovascular disease. In this paper, the calibration phase will be studied in order to distinguish between two emotional states: i) meditation and ii) stress condition. For this, the Heart Rate Variability (HRV) features are used as extracted from the RR interval and a support vectors machine (SVM) classifier deployed which resulted in 74% and 87% recognition accuracy based on HRV data for the recognition of the two emotional states, namely meditative, and, stressed, respectively.

The main objective is to prevent Cardio-Vascular Disease (CVD) in healthy people and to treat those who already suffer from it. Creating a reference database was our first step in this research project. The sensor choice was made based on doctors' recommendations. The work methodology was as follows: first validate the « objective data » issuing from the calibration state. Second, set up the automatic algorithm and detect automatically the patient's emotional states during the experimentation period (subjective data). Third analyse the physical activities correlated to the blood pressure and emotions. This study has involved the challenge of distinguishing the influence of stress versus relaxation on the Cardio-Vascular function and in particular on the risk of exacerbation of pre-existing Cardio-Vascular Disease.

* Tlija Amira, Email: tlija.amira@gmail.com

1. Introduction

The Internet of Things (IoT) represents a vision of ambient intelligent devices to support smart environments, lifestyles and hyper-connectivity based on the semantic interoperability of network-centric sensor devices as connected objects [1]. It represents different types of connection (Bluetooth, wireless, etc.) between several sensors within pre-established protocols. The connected objects reach several areas and fields. They can be found in the thermal energy sector, automated manufacturing, smart agriculture, intelligent buildings and healthcare sector. They are linked and connected in order to record, communicate and exchange data in real-time (Figure 1).

Chao Li [2] mentioned that by 2020 the number of connected sensors worldwide would reach 50 billion. It means that in less than 5 years the sensors number have almost doubled (in 2015 it was 25 billion objects). Combining these sensors with computer science tools like big data would increase the French gross domestic product (GDP) by 7% in 2025 [3]

Fig. 1. Data exchange and connected devices (Pixabay)

Hata [4] maintains that our world is mainly composed of three major fields: the environment and its influence on human beings, the technology that enables the sensing and perception of the environment and the physiological and emotional data of the human beings themselves, which in this study in multi-modal HealthCare Monitoring are connected through sensor devices by a complex system-of-systems that includes many operators and exploits the IoT architecture. When operating a healthcare system, the patient's environment must be managed ensuring their security and safety. The number of connected sensors and the necessity to record data in continuous and anonymous way, accurately using a self-synchronised data acquisition method constitute the paradigmatic challenges in this domain. The sensors measure various parameters and provide real-time data streams that need to be processed for data fusion and model building. Hata's health management framework [4] is based on a causality model as is this case study. Sensors measure the psycho-physiological states of the user as well as the relevant states of their living environment and the interaction between them. Emotions and human responses are closely related to hormone balance and the environment (lifestyle, work-style). In order to analyse emotion, the fact that feelings are personally interpreted must be taken into consideration. Some people cannot or will not identify their feelings and even if they do, it is still a subjective assessment. For example, a person might state: "I feel stressed" but physiologically it may that the person is just "tired", or may state "I feel relaxed" while the data indicates that they are stressed. The intercommunications between several areas in addition to the complication of administrating data and decision-making have to be taken into consideration.

2. Cardiovascular diseases

As mentioned in 2017, cardiovascular disease was one of the main causes of death in the world (according to the World Health Organisation). This affects 17 million people all over the world. CVD mainly concerns heart stroke and blood vessels. These diseases are not infections and cannot be passed from person to person. There are different types of CVD, the most common being coronary disease which causes one death in three in the United States every year (source centres for diseases control and prevention). In France, the main cause of death is due to CVD by 33% compared to 5% from digestive system diseases (source INSERM- INSEE). This kind of chronic disease is controlled by risk factors. Some of them cannot be changed like age, gender, family history. However, others can be moderated such as physical activities, obesity, high blood pressure, etc. by reducing those risks with healthier and better life style choices. This is the main objective of this study: to ensure better health care system using information technology. Chest pressure, heart pain, shortness of breath, nausea and feeling heartburn are the main symptoms of heart attack. These aspects are warnings that our body sends us of a dysfunction in our body.

Kernay [5] emphasised the importance of preventing, controlling and treating hypertension diseases, with high priority worldwide especially for developed countries. As figure 2 shows, being in developed or developing countries does not make a significant difference in hypertension risks. Being able to avoid hypertension, coronary disease, and peripheral arterial disease among other CVD is a worldwide healthcare system challenge. The European Cardiovascular Disease report for 2017 [7], shows that despite the age, sex and ethnicity difference between European citizens, CVD still the leading cause of death in Europe.

The aim is to prevent the disease if possible and to manage its treatment using data-analytic tools and connected sensors. Three main factors for CVD have been chosen with doctor's rrecommendations: Physical activities, regulation blood pressure and controlling emotions.

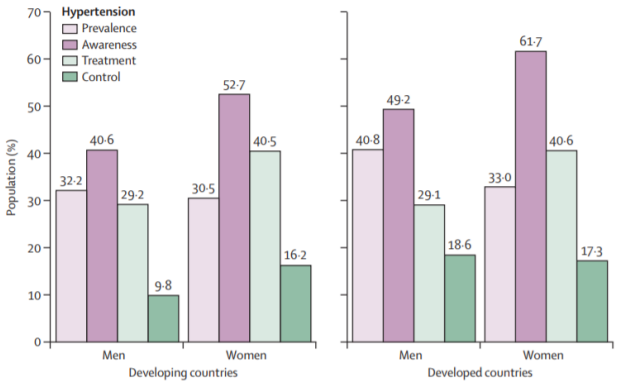


Fig. 2. Developed and developing countries: hypertension statistics (Source [6])

In this paper, a methodology will be presented to describe the main idea and objectives, define the experimentation protocol, explain the sensor choice and show the main results segmentation. However, the most challenging part in this research is controlling emotions for stress. For that, how emotion can be detected is explained based on a complex system of heart rate variability.

3. Heart rate variability

The heart rate variability measures the inter-beat interval. Analysing heart rate variability based on RR interval was the focus of the research. Much research has been conducted to show that there is a link between the emotional state, physiological and physic activities.

The HRV analysis can be divided into three areas: the long term (within 24hrs), the short term (about 5 min) and the ultra-short term (less than 5 min). Some research proposes the ultra-sort term from 1 min to 4 min [8]. Between the time domain, the frequency domain and the nonlinear metrics there are about 50 indicators for HRV. Monitoring people during daily routine life was the focus of this research. The calibration phase duration is about 10 minutes of meditation (calm respiration in order to calculate the heart rate frequency in a resting state) and 15 minutes of stress (stressful games like the Stroop colour test, mental arithmetic, etc.). The RR interval is being recorded with a frequency of one per second. In the literature, the HRV analysis has been analysed for a short-term duration and 24hr measurements. There is a total absent of studies conducted on ultra-short term. The purpose of this research is to bring out several indicators relevant to identifying each emotional state during ultra-short-term testing.

The Autonomic Nervous System (ANS) is the part of the nervous system in our body that regulates the function of human organs and the cardiac muscles and glands. It is the vital system for controlling many phenomes in our body: such as temperature changes sending extra blood to some organs, causing a slower heartbeat, etc. This system is divided in two main areas: the parasympathetic system and the sympathetic system. They serve the same organs but have opposite roles. While one excites our body and supports fight or flight reactions to the environment, the other calms it down. When feelings of anxiety, panic, fear, happiness, etc. are experienced, these are initiated by our sympathetic system thus preparing the body for physical activities and emotional responses. In other words, "stress and emotions" are evoked mainly by the sympathetic system while "relaxation and calm responses" are evoked by the parasympathetic system. The analysis of emotions is the focus of this work on stress based on HRV that was calculated from the RR interval extracted from the heart rate monitor belt. The data recording is 1s for each observation. There is not a unique indicator that can categorise the HRV. This complex system takes into consideration the paradoxical behaviour between the parasympathetic system and the sympathetic system as well as the respiratory mechanism. In much research, the LF/HF ratio (Low Frequency/High Frequency) has been considered reliable in presenting sympathovagal balance. Wilhelm [9] confirms that using this ratio for a two-dimensional representation framework can help in stress recognition. However, using only this indicator is not enough for this research thus analysing daily data including different kinds of emotion, such as fear, panic, happiness, stress, satisfaction, etc. This paper presents the potential benefit of using a Support Vector Machine (SVM) to identify mental stress based on heart rate variability. The complexity of heart rate indicates the flexibility of the Autonomic Nervous System (ANS) and sympathovagal modulation effects on HR variability. Several HRV indicators were taken into consideration from the time and frequency domains to Poincaré and statistical analysis. The objective was to distinguish between the effects of sympathovagal fluctuations and emotion on HR variability.

Emotions and the way they are expressed are different from one person to another. Many studies were conducted into this area, and most of them show that having a positive attitude, surrounded by positive feelings, helps people to live longer compared to those who are always feeling angry, afraid and stressed. In 1977, the American psychologist Caroll Ellis Izard classified emotions into ten categories; joy, surprise, sadness, anger, disgust, contempt, shame, fear, guilt, interest or excitement. As cited, stress is not technically an emotion. It is a feeling that based on a reaction or a stimulus. A recent study in 2018 [10], based on the state-of-the-art of 37 publications showed that heart rate variability is a reliable indicator for stress recognition.

4. The proposed work and other related works

This is not the first work undertaken on stress and HRV as mentioned in the related work in the paper. But the novel contribution is that this work will be the first to separate different kinds of emotion during daily life. No related work has been conducted that mentioned how to distinguish between happiness, being angry, stressed or feeling relaxed. No related research undertook daily monitoring for patient/participants.

There have been no previous studies deploying a similar methodology of longitudinal real-life study of the psycho-physiological correlates of evolution and exacerbation of cardio-vascular conditions. This study was initiated responsive to earlier mainly clinically controlled studies [11], [12], [7] which involved data capture in the context of a few specifically detectable activities (such as preparing food, climbing stairs, using a smartphone, etc.). As of 2018 some work has been mainly concerned with feature extraction [9]; others have used various approaches for stress detection e.g. a survey approach [10], saliva and cortisol level [11]. Others have used controlled studies to evaluate the impact of fast food consumption [12] lifestyle in terms of cardio-vascular disease risk and some have involved data capture in an IOT-enabled controlled smart home context [13]. The approach followed here is distinguished by being based on an ambulant environment (rather than being clinically controlled), connected-devices-enabled continuous data measurement including contemporaneous self-expression recordings of the participant's perceived feelings and mood. The project objective is to extract information and correlations from different types of data (corporal and sentimental) and to bring up analysis that will present the impact of these variables and features on cardiovascular diseases in terms of prevention and treatment. The experiments started with data acquisition involving healthy individuals paving the way for addressing sensor integration, data acquisition, storage and integration issues and evaluating the experimental set-up for the future. This paper explains the project methodology, experiments

already conducted and those planned to be carried out within INSEAD-Sorbonne Université. Firstly, the sensor choice was explained, secondly, the concept and objectives were described and demonstrated and the questionnaires were tested and validated. Finally, the experimentation and analysis tools were defined. sensors currently available on the market can support the measurement of three categories of data regarding the cardio-vascular function; namely: i) the physiological parameters such as activity level, ii) arterial pressure, iii) stress level. The research purpose is to build an automatic system detection capable of recognising different types of emotion and to categorising feelings and mood according to heart beat, arterial pressure among several other variables. Moreover, the work's contribution is to ensure monitoring participants while they are wearing sensors during their daily life. Despite the fact that not all the stress-creating factors are manageable, the conducted experience gave rise to reliable data that illustrate participants' spontaneous feelings and mood. As the experimentation has been conducted over a long period, the part that concerns data storage and management had to be handled securely.

5. Materials and methods

5.1. Wearables sensors

Before launching this research, a literature review was conducted . The idea was to study the types of existing sensors on the market and to the research carried out already [8]. The sensor choice was based on three hypotheses that are explained as follows:

1. Blood Pressure or the blood circulation is highly correlated with cardio vascular diseases[14, 15] and thus identifying the flow of blood circulation with a connected sensor remains an important part to maintain a stable blood pressure. Hastie and Gluszek [16, 17] mentioned that it is essential to measure heart rate variability in order to manage and control the risk of having a cardiovascular disease.
2. Nutritional attitude and physical activities are two important factors that people should take into consideration in order to prevent the apparition of cardiovascular diseases [18].
3. Feeling overwhelmed, anxious, stressed among several other feelings, are factors that can be the main cause for the appearance of chronic diseases. These cognitive and emotional factors, have not yet been measured in an ambulatory approach [13, 19, 20].

To satisfy the hypothesis mentioned above and to set up this experiment, 3 connected objects were selected. They will measure the physical activity, the blood pressure and heart rate variability. So each participant was given a connected tensiometer, a heart rate monitor belt and a smart watch.

Tensiometer

Following a study of user acceptance, reliability and wearability criteria, the tensiometer «Rest» AD761f was selected as one of the sensors to be integrated within the body area sensor network (see Fig.3). Tensiometers of this brand are among the first to have passed the European Society of Hypertension (ESH) protocol [21], [22]. This device enables the detection of a person's rest condition using the Hemodynamic Stability Determination technique (HSD).

Fig. 3. Rossmax tensiometer

The smartwatch ActiGraph

Dinseh [23,24], states that the actigraph is one of the most reliable widely used accelometers in scientific research. This smart watch is an activity monitor (Figure 4) tested, tried and approved by many researchers and practitioners [25, 26] as a reliable sensor giving accurate data. As the trial was over a long period and the objective was to monitor people during their daily life, many limitations were imposed on the sensor choice. It had to be small, discreet, user friendly, giving reliable data. As the sensor placement affects the degree of data accuracy, all participants were asked to carry this connected device in the same position on their waist (Figure 4)

Both raw and calibrated data were extracted at a range of sampling frequencies (ranging from 30 Hz to 100 Hz).

Fig. 4. Actigraph sensor worn

The heart rate monitoring belt: Polar H7

For emotional state monitoring and measurement specifically stress, a state-of-the-art analysis of the emotional states sensing technology was performed as well as a study of the readymade solutions for this which could be mass-scalable. To quantify emotions and feeling many tools and techniques are available specially to measure stress among other feelings. This mental pressure and feeling is one of the main factors that accentuates the appearance of CVD among chronic diseases [19] [20] [55].

Marek and Hamilton [33, 34] specified measuring HRV as a reliable technique for categorisation of several kinds of emotions. For this reason participants were equipped with the heart rate monitor belt to measure HRV based on the RR interval (figure 5)

Fig. 5. The Heart Rate Monitor Belt

5.2. Methodology

The methodology was based on the aim of this work, to treat people with CVD and to prevent the apparition of chronic diseases especially cardiovascular ones. This project methodology was centred on studies that explain the association between the emotional states and heart rate variability [35, 36].

One of the important emotion types to be monitored is stress which can lead to CVD, specifically chronic stress increasing blood pressure and exacerbating CVD; the physiological and psychological reactions to stress can manifest as both anxiety and depression. The stress monitored was real life stress and not induced or simulated for experimental purposes [37]. Vaping and smoking, being overweight or obese, the absence of physical activities, being anxious and stressed are factors that will put healthy participants at risk. The trial duration was 15 days in which participants wore the three sensors according to the user manual. No restriction was imposed. Participants were not asked to change their daily routine nor to be more active. This study is distinguished by the fact that the key psycho-physiological data relevant to cardio-vascular function was captured continuously over a long period to support offline analysis, rather than on a one–off diagnostic basis. This trial was validated in three steps. First, the experts decided the inclusion and the exclusion criteria to select different participants. Second, three online surveys were tested on the UTC team. According to their feedback more questions were added. Third, the procedure was validated and all the experimentation protocol was validated by the INSEAD ethical committee.

5.3. The experimental protocol

This project was established in collaboration with the Institut Europeen d'Administration des Affaires (INSEAD). The experimentation protocol took the welfare of the participants into consideration. The main research goal is to measure some variables via the sensors and devices and to extract features from this data. The main work needs to be done on patients with CVD. Nevertheless, the study with the INSEAD lab was performed on healthy people. A database reference was necessary and so the trial was launched with participants with no chronic diseases who were closest to the sociodemographic criteria of France.

Before launching this experiment, two main exclusion criteria were imposed: having a chronic disease and not reaching the age of majority (18-19 years old). People would wear the sensors every day during the trial. Physiological and emotional parameters will be measured continuously.

Candidates were assigned coded identities, which were transmitted once the trial started. The protocol ID is ID February2018/1. It was approved by the Ethical Committee in INSEAD. All data are anonymously collected and treated.

A quantitative study was performed before starting the recordings. First doctors from different specialities (neurologist, cardiologist and generalist) were questioned. This enabled the categorisation of the three main areas that will affect chronic diseases generally and CVD specifically. Next a focus group was set up with the UTC lab in order make sure that all the questions are clear and easy to understand (as few questions have medical terms). Age, antecedent diseases, physical activity, eating habits and emotion were the main features that doctors proposed to focus on. An emotional eating a validated questionnaire concerning emotion and eating was conducted inspired by Garaulet et al. [38].

The three main questionnaires were implemented on Qualtrics (within INSEAD Sorbonne université). Each participant had to enter his/her own code to express how the study was conducted. The main three surveys are as follow:

1. The pre-selection questionnaire: Selecting people according to the inclusion criteria
2. Before experiment questionnaire: Analysing participants' physical activity, eating habits, the general assessment of the candidate's health (smoking, alcohol, obesity ...).
3. After experimentation questionnaire:the main purpose is to study the candidate's satisfaction with the experience, focus on participant's recommendation, find out if the experiment helped candidate to change his/her dietary habits and becoming an active person.

a) Procedure

The trial duration is 2 weeks; during which candidates must wear three connected objects. The heart rate monitor belt should be worn at least 2-3 hours per day. The tensiometer for morning and evening measurements and the activity tracker all day long.

After selection of each participant, a meeting took place with them in order to explain experimentation's protocol. Participants were informed that we created a forum free place to ask questions and exchanging with each other and giving their feedback about the experience. They were given a user manual re the tools, how and when to wear sensors. They were also given a notepad in which to write down their feelings, physical activities and emotions while they are wearing the sensors.

They were also told that the experiment was not to be regarded as an alternative to any medical or other treatment and informed about the FORUM, a free and anonymous point of exchange of ideas to be used to analyse participants' responses [39], enabling participants to write their feelings anonymously, their issues as well as their remarks and questions. Before starting the experience, candidates had to read and sign the protocol. Once the study is over, they have to bring back the sensors to INSEAD and so getting a financial compensation for their effort during this trial (Figure 6)

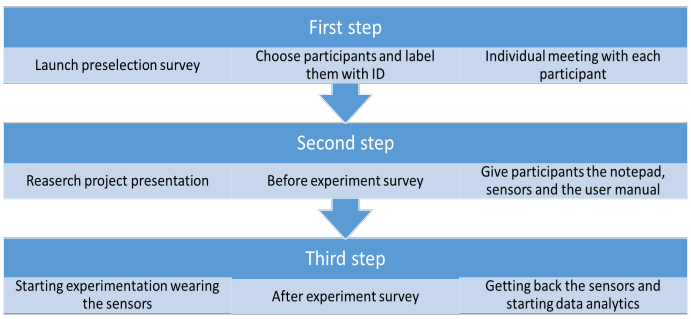


Fig. 6. The description of the experimentation protocol

b) Subjects

Before recruitment began, the intention was to have a population similar to the French population described by the INSEE; 23 men and 22 women. However, recruiting people depends on their flexibility, motivation, availability and eligibility. The trial duration was 4 to 5 months during which 53 participants took part wearing the devices for 15 days each. Candidates were from different socio-professional categories, different levels of study and age category. In this section, some statistical results of the population are presented and participant feedback.

As the graph shows (Figure 7) almost 40% of our population are healthy people with no cardiovascular disease who are aged over 40 years. Recruiting men for this study was not a simple task.

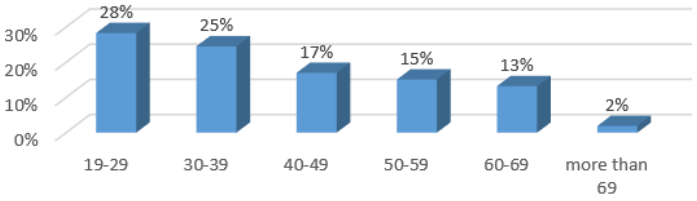


Fig. 7. Age Categories

The final population is composed of ¾ women and ¼ men. Obtaining certain variables during sensor initialisation enabled the comparison of the participants' measurements. The Body Mass Index (BMI) is calculated as follows: BMI = mass / height^2. The interpretation is as follows:

- If BMI <= 18.5 : the participant is in underweight area
- 18.5 <BMI <= 25 : the participant is in a normal weight area
- 25 <BMI <= 30 : the participant is overweight
- 30 <BMI <= 35 : the participant is in a moderate obesity area
- 35 <BMI <= 40 : the participant is in a severe obesity area
- 40 <IMC : the participant is in an obesity morbid area

57% of participants have a normal weight with only 4% having morbid obesity (see Figure 8). 31% is the wear time validation for less than 40% of participants. This shows that participants were motivated to carry out the experiment

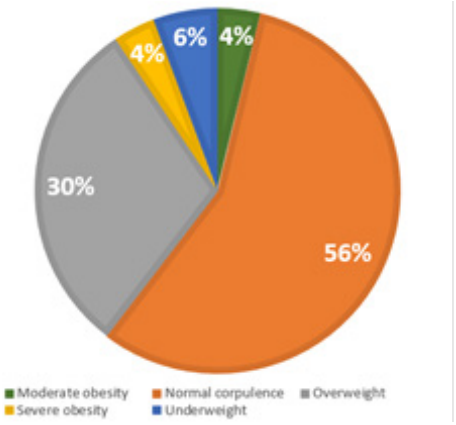


Fig. 8. Obesity Level

6. HRV and calibration phase

Heart rate variability is highly used for emotions analysis like anxiety, fear and specially stress. In fact, the HRV is described by several parameters and biomarkers. There are features that are calculated in the statistical domain, time serial domain and others are in the frequency domain. One powerful biomarkers for stress detection based on HRV analysis is the LF/HF ratio (Low frequency/ High frequency).

The interpretation is based on two hypotheses: the ratio LF/HF must be interpreted. If its value is less than one it shows that the person is feeling relaxed, sleeping or just resting. Otherwise it means that the participant is carrying out a physical activity, running or feeling stressed

The first HRV analysis of recordings on Kubios, showed the confirmation or the two hypotheses mentioned above (Figure 9 and 10)

Fig. 9. Kubios HRV analysis that shows a participant doing a physical activity

Fig. 10. Participant in rest condition (yoga class)

Before starting this experiment with the collaboration with INSEAD Sorbonne Université, a « blind test » was set up with the UTC member lab. The idea was to make sure that the equipment and sensors are connected and that they were friendly user.

While analysing the HRV based on participants, notes (on the notepad) it was deduced that sometimes there is no correlation between what is written on the notepad and the HRV analysis. In fact, emotions are our personal interpretation. For example,

participant A can write on the notepad 'feeling stressed' and while analysing data of participant A we found that he was not stressed. In addition, the fact that people sometimes have difficulty identifying their feelings has to be taken into consideration. It was therefore decided that 'reliable data' was needed which meant recording data while observing what is actually happening. This was called that the 'calibration state'. Games were developed to stress people. Before starting the experimentation, participant had to spend 20 min of recording 'reliable data' starting with 10 min of deep respiration and meditation. After that 10 minutes of playing stressful games that are reported as follows: Starting with 3 minutes of a mathematical mental exercise, next playing the ball game for 3 minutes and finishing with the Stroop Colour Word Test, the cognitive stressful test. These games were developed using Visual Studio and Scratch (Fig. 11). Despite the fact that the game's idea is validated in advanced research [40], it was decided to develop them so they can be user friendly in an interactive way and adjust the speed games, another helpful factor for boosting the participant's stress. Before launching these games, they were trialled on 13 users from the UTC Lab and the graphical interface was adjusted according to their feedback and remarks.

Fig. 11. Interactives games

7. Support Vector machine

To analyse the HRV there are distinct techniques that enabled the extraction of several features. The support vector machine is often used in statistical machine learning theory to solve classification issues in several areas in the medical field. Lanata [41] used this algorithm to identify horses' response to human fear and happiness. Mirhoseini [42] used it to detect early cardiac death while this pattern recognition classifier was also used for health care applications based on Heart Rate Variability [45-50]

The main use of this supervised algorithm is to distinguish between two classes (or more). Having a set of points in a feature space, each point is associated to a label. The dispersion of a group of points with the description is made up of a straight line or a hyperplane (in the context of two dimensions).

Nevertheless, the problem of machine learning only functions if the data has a linear structure, which is not the case here where the SVM with RBF Kernel (Radial Basis Function Kernel) was used to overcome this.

8. Results

Fred Shaffer and Roman Baevsky [43] [44] mentioned all the state of the art of research that has used the ECG signal and RR interval to interpret the HRV. The results shown in this section, concern the calibration phase. An Ultra short-term period (less than 5 min of analysis). Corresponding features were extracted and some statistical variables were added. A feature was created [24] for arrhythmia detection (if the HTI <= 20.42 and RMSSD<= 0.068 then the heart rhythm is normal, and if the HTI> 20.42 then there is a presence of arrhythmia). The 'HRV Tool' implemented by Marcus Vollmer [53] on Matlab was used; in order to extract some features from the RR signal. During this calibration phase 34 participants were chosen and the data divided into two categories: stress label and relaxation label. Work was chosen with a signal processing 'windows of 90s with recovering of 50% [45]. The 24 selected features are described in the table 1.

Table 1. Different features

Domain	Features
Time Domain	SDSD, PNN50, NN50, RMSSD, HRV triangular index (HTI) Mean RR, Standard deviation, coefficient of variation, rrHRV and TINN[44] , [45]
Frequency Domain	VLF = Very Low Frequency (0.0033-0.04 Hz), LF= Low Frequency (0.04-0.15 Hz), HF = High Frequency (0.15-0.4 Hz), pLF, pHF and LF/HF ratio[46],[47]
Statistics analysis	KURTO, Skeww, Interquartile, Arrythmia (that detect the arrhythmia corresponding to [44])
Poincare plot	SD1, SD2 and SD1SD2ratio, approximate entropy (ApEn)[49] ,[50]

The data was divided into two datasets: 80% for training data and 20% for test data. We made a k fold (k= 5) so the final obtained value is an average of the 5 iteration inside the SVM model. The table below shows the main results with 74.44% of good recognition rate:

Table 2. Confusion Matrix

	Class 1	Class 2	% good recognition rate
Class 1	48	38	55.81 %
Class 2	17	115	87.12 %

9. Discussion and conclusion

Our study mainly concerned monitoring patients with chronic diseases at home and controlling heart rate variability with wearable sensors. The experimental protocol was conducted with 50 participants who were equipped with 3 connected objects which. they had to wear for 15 days and to record, on the notepad provided, what they were doing (activity level) and how they were feeling throughout the day. Before launching this experiment and in order to ensure the accuracy of the acquired data, each participant had to undergo a « calibration phase ». In this paper the

methodology and the results obtained from analysing the calibrated data are presented. The SVM on the RR measurements were applied to interpret the heart rate variability based on the activity and stress levels during any period of observation.

34 RR interval signals, including two different beat types (stress and meditation) were selected. 24 features were selected and the support vector machine algorithm was applied. The HRV signal extraction from each RR interval showed over 74% of accurate recognition. The optimisation of the use of features and thus the reduction of the error rate will be the aim of future work. The algorithm has a recognition accuracy of over 87% for class 2 that indicates the stress induced from playing games. The population was divided by age and gender in order to obtain a more efficient classification [54].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was funded within the framework of E-BIOMED Chair IUIS (Institut Universitaire d'Ingénierie en Santé). Experiment will be conducted in collaboration with the Centre Multidisciplinaire des Sciences Comportementales Sorbonne Universités–Institut Européen d'Administration des Affaires (INSEAD).

References

[1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks, vol. 10, no. 7, pp. 1497–1516, 2012.

[2] C. Li, X. Hu, and L. Zhang, "The IoT-based heart disease monitoring system for pervasive healthcare service," Procedia Comput. Sci., vol. 112, pp. 2328–2334, 2017.

[3] I. Montaigne, "Big data et objets connectés Faire de la France un champion," Rapport, p. +200, 2015.

[4] Y. Hata, S. Kobashi, and H. Nakajima, "Human Health Care System of Systems," IEEE Syst. J., vol. 3, no. 2, pp. 231–238, 2009.

[5] Y. G. Smeyers and M. Villa, "A study of the internal dynamics of inverting trimethylamine by means of the non-rigid group theory," Mol. Phys., vol. 101, no. 12, pp. 1795–1803, 2003.

[6] K. B. Tibazarwa and A. A. Damasceno, "Hypertension in developing countries," Can. J. Cardiol., vol. 30, no. 5, pp. 527–533, 2014.

[7] Wilkins E. et al., "European Cardiovascular Disease Statistics 2017 edition," Eur. Hear. Network, Brussels, p. 192, 2017.

[8] W. von Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, and D. P. Mandic, "Resolving ambiguities in the LF/HF ratio: LF-HF scatter plots for the categorization of mental and physical stress from HRV," *Front. Physiol.*, vol. 8, no. JUN, pp. 1–12, 2017.

[9] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, and B. H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," Psychiatry Investig., vol. 15, no. 3, pp. 235–245, 2018.

[10] C. Zhu et W. Sheng, « Multi-sensor fusion for human daily activity recognition in robot-assisted living », in Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, 2009, p. 303–304.

[11] L. Bao et S. S. Intille, « Activity recognition from user-annotated acceleration data », in Pervasive computing, Springer, 2004, p. 1–17.

[12] A. J. Brush, A. K. Karlson, J. Scott, R. Sarin, A. Jacobs, B. Bond, O. Murillo, G. Hunt, M. Sinclair, K. Hammil, et others, « User experiences with activity-based navigation on mobile devices », in Proceedings of the 12th international conference on Human computer interaction with mobile devices and services, 2010, p. 73–82.

[13] D. Huysmans et al., "Unsupervised Learning for Mental Stress Detection - Exploration of Self-organizing Maps," Proc. 11th Int. Jt. Conf. Biomed. Eng. Syst. Technol., no. January, pp. 26–35, 2018.

[14] Z. Jinbiao, L. Mao, and S. Oceanic, "- 8. 6," vol. 10, no. 4, pp. 609–611, 1991.

[15] C. Höcht, "Blood Pressure Variability: Prognostic Value and Therapeutic Implications," ISRN Hypertens., vol. 2013, p. 16, 2013.

[16] C. E. Hastie, P. Jeemon, H. Coleman, L. McCallum, R. Patel, J. Dawson, W. Sloan, P. Meredith, G. C. Jones, S. Muir, M. Walters, A. F. Dominiczak, D. Morrison, G. T. McInnes, and S. Padmanabhan, "Long-term and ultra long-term blood pressure variability during follow-up and mortality in 14,522 patients with hypertension.," Hypertension, vol. 62, no. 4, pp. 698–705, 2013.

[17] J. Głuszek and T. Kosicka, "Visit-to-visit blood pressure variability and cardiovascular and kidney diseases," Nadcisnienie Tetnicze, vol. 20, no. 1, pp. 26–31, 2016.

[18] M. et al. Poelman, "Relations between the residential fast-food environment and the individual risk of cardiovascular diseases in the Netherlands: a nationwide follow-up study," Eur. J. Prev. Cardiol., no. April, 2018.

[19] P. Krolak-Salmon, "La reconnaissance des émotions dans les maladies neurodégénératives," Rev. Med. Interne, vol. 32, no. 12, pp. 721–723, 2011.

[20] S. Montel and C. Bungener, "Les troubles de l'humeur et des émotions dans la sclérose en plaques : une revue de la littérature," Rev. Neurol. (Paris)., vol. 163, no. 1, pp. 27–37, 2007.

[21] Tasker F, de Greff A, Liu B, Shennan AH "Validation of a non-mercury ascultatory device according to the European Society of Hypertension protocol: Rosssmax Mandaus II. Blood Press. Monit".14, pp.121–4, 2009.

[22] Zhang, L., Kang, Y. Y., Zeng, W. F., Li, Y., & Wang, J. G, "Validation of the Rossmax CF175 upper-arm blood pressure monitor for home blood pressure monitoring according to the European Society of Hypertension International Protocol revision 2010", Blood pressure monitoring. Rev. 20(2), pp. 104-107, 2015.

[23] P. Under and T. H. E. Hood, "NIH Public Access," vol. 44, pp. 1–6, 2013.

[24] A. Manuscript, A. Vertical, and A. Counts, "NIH Public Access," vol. 33, no. 2, pp. 219–230, 2013.

[25] Coker-Bolt, P., Downey, R. J., Connolly, J., Hoover, R., Shelton, D., & Seo, N. J. "Exploring the feasibility and use of accelerometers before, during, and after a camp-based CIMT program for children with cerebral palsy". Journal of pediatric rehabilitation medicine, 10(1), pp. 27-36, 2017.

[26] S. R. Dandu, M. M. Engelhard, A. Qureshi, J. Gong, J. C. Lach, M. Brandt-Pearce, and M. D. Goldman, "Understanding the Physiological Significance of Four Inertial Gait Features in Multiple Sclerosis," IEEE J. Biomed. Heal. Informatics, vol. 22, no. 1, pp. 40–46, 2018.

[27] Williams, R. A., Hagerty, B. M., & Brooks, G "Trier Social Stress Test: A method for use in nursing research. Nursing research", 53(4), 277-280, 2004.

[28] R. Schollmeier and T. Universitat, "A Definition of," vol. 15, no. 1, pp. 101–102, 2002.

[29] M. Gröschi, "Données actuelle sur l'analyse hormonale salivaire. Current status of salivary hormone analysis," Ann Biol Clin, vol. 67, no. 5, pp. 493–504, 2009.

[30] D. Nishimura, "Graphically speaking," Science (80-.)., vol. 283, no. 5405, p. 1134, 1999.

[31] M. R. Ebben, V. Kurbatov, and C. P. Pollak, "Moderating laboratory adaptation with the use of a heart-rate variability biofeedback device (StressEraser®)," Appl. Psychophysiol. Biofeedback, vol. 34, no. 4, pp. 245–249, 2009.

[32] M. E. G. H. O'Donnell C., "Detection and management," EmbaseTransfusion Med., no. October 2017, pp. 64–69.

[33] M. Marek, "Guidelines Heart rate variability," Eur. Heart J., vol. 17, no. March, pp. 354–381, 1996.

[34] J. L. Hamilton and L. B. Alloy, "Atypical reactivity of heart rate variability to stress and depression across development: Systematic Review of the literature and directions for future research," Clin. Psychol. Rev., vol. 50, pp. 67–79, 2016.

[35] M. R. Gunnar, D. Ph, A. Herrera, C. E. Hostinar, and B. Sc, "Stress et développement précoce du cerveau," pp. 1–8, 2009.

[36] J. Ham, D. Cho, J. Oh, and B. Lee, "Discrimination of multiple stress levels in virtual reality environments using heart rate variability," 2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 3989–3992, 2017.

[37] M. Buckert, J. Oechssler, and C. Schwieren, "Imitation under stress," J. Econ. Behav. Organ., vol. 139, pp. 252–266, 2017.

[38] M. Garaulet, M. Canteras, E. Morales, G. López-Guimera, D. Sánchez-Carracedo, and M. D. Corbalán-Tutau, "Validación de un cuestionario de comedores emocionales para uso en casos de obesidad; cuestionario de comedor emocional (CCE)," Nutr. Hosp., vol. 27, no. 2, pp. 645–651, 2012.

[39] B. Liu, "Sentiment Analysis and Opinion Mining," no. May, pp. 1–108, 2012.

[40] D. Huysmans, E. Smets, W. De Raedt, C. Van Hoof, K. Bogaerts, I. Van Diest, and D. Helic, "Unsupervised Learning for Mental Stress Detection - Exploration of Self-organizing Maps," Proc. 11th Int. Jt. Conf. Biomed. Eng. Syst. Technol., no. January, pp. 26–35, 2018.
[41] Lanata, A., Nardelli, M., Valenza, G., Baragli, P., D'Aniello, B., Alterisio, A., ... & Scilingo, E. P. (2018, July). A Case for the Interspecies Transfer of Emotions: A Preliminary Investigation on How Humans Odors Modify Reactions of the Autonomic Nervous System in Horses. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 522-525). IEEE.
[42] Mirhoseini, S. R., JahedMotlagh, M. R., & Pooyan, M. (2016). Improve accuracy of early detection sudden cardiac deaths (SCD) using decision forest and SVM. In International Conference on Robotics and Artificial Intelligence (ICRAI), USA Google Scholar.
[43] Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in public health, 5, 258(over view of heart rate)
[44] Jovic, A., & Bogunovic, N. (2011). Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. Artificial intelligence in medicine, 51(3), 175-186.
[45] Guo, H. W., Huang, Y. S., Lin, C. H., Chien, J. C., Haraikawa, K., & Shieh, J. S. (2016, October). Heart Rate Variability Signal Features for Emotion Recognition by Using Principal Component Analysis and Support Vectors Machine. In 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 274-277). IEEE
[46] Salahuddin, L., Cho, J., Jeong, M. G., & Kim, D. (2007, August). Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE (pp. 4656-4659). IEEE.
[47] Park, D., Lee, M., Park, S., Seong, J. K., & Youn, I. (2018). Determination of Optimal Heart Rate Variability Features Based on SVM-Recursive Feature Elimination for Cumulative Stress Monitoring Using ECG Sensor. Sensors, 18(7), 2387.
[48] Venkatesan, C., Karthigaikumar, P., Paul, A., Satheeskumaran, S., & Kumar, R. (2018). ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications. IEEE Access, 6, 9767-9773.
[49] Ashtiyani, M., Lavasani, S. N., Alvar, A. A., & Deevband, M. R. (2018). Heart Rate Variability Classification using Support Vector Machine and Genetic Algorithm. Journal of Biomedical Physics and Engineering, 8(4 Dec).
[50] Lyle, J. V., Charlton, P. H., Bonet-Luz, E., Chaffey, G., Christie, M., Nandi, M., & Aston, P. J. (2017). Beyond HRV: analysis of ECG signals using attractor reconstruction. Computing, 44, 1.
[51] Lee, H. G., Noh, K. Y., & Ryu, K. H. (2008, May). A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on (Vol. 1, pp. 200-206). IEEE.
[52] Baevsky, R. M., & Chernikova, A. G. (2017). Heart rate variability analysis: physiological foundations and main methods. Cardiometry, (10).
[53] Vollmer, M. (2015, September). A robust, simple and reliable measure of heart rate variability using relative RR intervals. In Computing in Cardiology Conference (CinC), 2015 (pp. 609-612). IEEE.
[54] BLK Antelmi, I., De Paula, R. S., Shinzato, A. R., Peres, C. A., Mansur, A. J., & Grupi, C. J. (2004). Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease. The American journal of cardiology, 93(3), 381-385.
[55] Amira, T., Dan, I., Az-eddine, B., Ngo, H. H., Said, G., & Katarzyna, W. W. (2018, June). Monitoring chronic disease at home using connected devices. In *2018 13th Annual Conference on System of Systems Engineering (SoSE)* (pp. 400-407). IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 40-44 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Proposed Architecture for Parallel HPC-based Resource Management System for Big Data Applications

Waleed Al Shehri*, Maher Khemakhem, Abdullah Basuhail, Fathy E. Eassa

Department of Computer Science, King Abdul-Aziz University, Jeddah, KSA

ARTICLE INFO

Article history:
Received: 02 October, 2018
Accepted: 11 January, 2019
Online : 20 January, 2019

Keywords:
High-Performance Computing
Big Data
Load-Balancing
Data-Locality
Resource Management
Parallel Programming models
Power Consumption

ABSTRACT

Big data can be considered to be at the forefront of the present and future research activities. The volume of data needing to be processed is growing dramatically in both velocity and variety. In response, many big data technologies have emerged to tackle the challenges of collecting, processing and storing such large-scale datasets. High-performance computing (HPC) is a technology that is used to perform computations as fast as possible. This is achieved by integrating heterogeneous hardware and crafting software and algorithms to exploit the parallelism provided by HPC. The performance capabilities afforded by HPC have made it an attractive environment for supporting scientific workflows and big data computing. This has led to a convergence of the HPC and big data fields.

However, big data applications usually do not fully exploit the performance available in HPC clusters. This is so due to such applications being written in high-level programming languages and do not provide support for exploiting parallelism as do other parallel programming models.

The objective of this research paper is to enhance the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. This can be achieved by building a parallel HPC-based Resource Management System to exploit the capabilities of HPC resources efficiently.

1. Introduction

The amount of data produced in the scientific and commercial fields is growing dramatically. Correspondingly, big data technologies, such as Hadoop and Spark, have emerged to tackle the challenges of collecting, processing, and storing such large-scale data.

There are different opinions on the definition of big data resulting from different concerns and technologies. One definition applies to datasets that cannot be realized, managed and analyzed with traditional IT software. This definition reflects two connotations: data volume that is growing and changing continuously; and, this growing volume is different from one big data application to another [1]. A more specific definition based on the multi-V model by Gartner in 2012: ‘‘Big Data are high-volume, high-velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization’’ [2].

While the focus of big data applications is on handling enormous datasets, high-performance computing (HPC) focuses on performing computations as fast as possible. This is achieved by integrating heterogeneous hardware and crafting software and algorithms to exploit the parallelism provided by HPC [3]. The performance capabilities afforded by HPC have made it an attractive environment for supporting scientific workflows and big data computing. This has led to a convergence of the HPC and big data fields.

Unfortunately, there is usually a performance issue when running big data applications on HPC clusters because such applications are written in high-level programming languages. Such languages may be lacking in terms of performance and may not encourage or support writing highly parallel programs in contrast to some parallel programming models like Message Passing Interface (MPI) [4]. Furthermore, these platforms are designed as a distributed architecture, which differs from the architecture of HPC clusters [5].

* Waleed Al Shehri, : Email: waleed.ab2@gmail.com

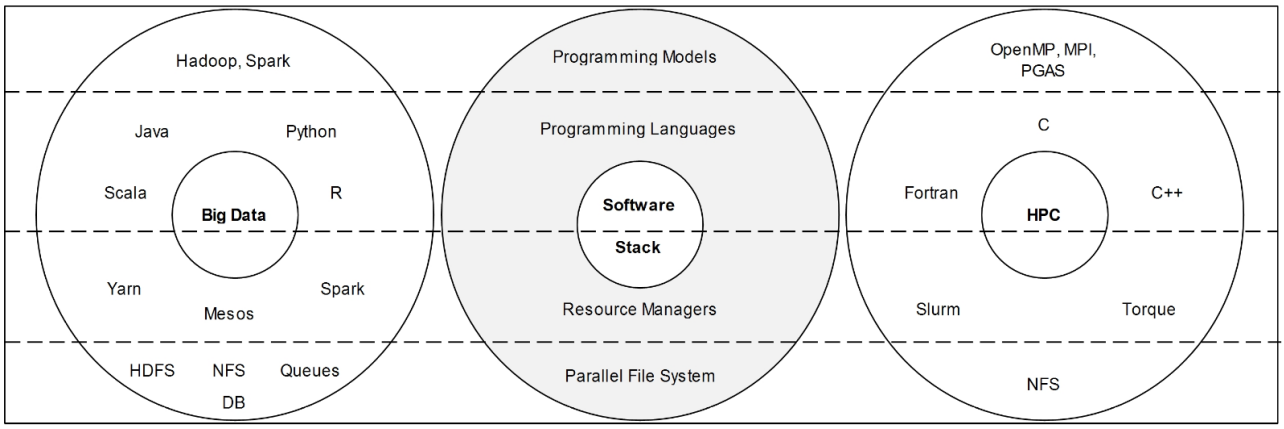


Figure 1. HPC and Big Data Software Stacks.

Additionally, the large volume of big data may hinder parallel programming models such as Message Passing Interface (MPI), Open Multi-Processing (OpenMP) and accelerator models (CUDA, OpenACC, OpenCL) from supporting high levels of parallelism [1].

Furthermore, resources allocation in HPC is one of the prime challenges, especially since HPC and big data paradigms has a different software stack [6] as shown in (**Figure 2**):

2. Related Work

The related work can be organized based on the different aspects required to fulfill the architecture requirements of this research. The job scheduler concept will be highlighted by considering its features and functionality. Moreover, different comparative studies will be introduced covering big data programming models and parallel programming models to establish the performance gap between them. Other works are presented to show how the performance of these programming models can be enhanced. Finally, research involving data locality approaches and decomposition mechanisms covering the same context of this research is reviewed.

A job scheduler can play an essential role in modern big data platforms and HPC systems. It manages different compute jobs related to different users on homogenous or heterogeneous computational resources. It can have different names that reflect the same mechanism such as scheduler, resource manager, resource management system (RMS), and a distributed resource management system (D-RMS) [7]. Despite significant growth in terms of heterogeneity of resources and job complexity and diversity, job schedulers still have the main core function of job queuing, scheduling and resource allocation, and resource management [8][9]. In [7], many features are analyzed of the most popular HPC and big data schedulers including Slurm, Son of Grid Engine, Mesos, and Hadoop YARN.

Additionally, there are two primary job types: job arrays and parallel jobs. In job arrays, multiple independent processes for a single job identifier can be run with different parameters for each process. In parallel jobs, it is possible to launch each of the processes simultaneously, allowing communication between them during the computation. While HPC schedulers support both types, big data schedulers can support only job arrays. Furthermore, there are many important features of HPC schedulers, generally not available with big data schedulers, such as job chunking, gang scheduling, network aware scheduling and power-aware scheduling.

Big data jobs are usually considered to be network-bound regarding a large amount of data movement between different nodes among clusters. In [10], traffic forecasting and job-aware priority scheduling for big data processing is proposed by considering the dependencies of the flows. The network traffic for flows of the same job is forecasted via run-time monitoring, then a unique priority for each job is assigned by tagging every packet in the job. Finally, it uses a FIFO order for scheduling flows of the same priority.

In [11], a new backfilling algorithm, known as fattened backfilling, is proposed to provide more efficient backfilling scheduling. In this algorithm, short jobs can be moved forward if they do not delay the first job in the queue. A Resource and Job Management System (RJMS) based on a prolog/epilog mechanism has been proposed in [12]. It allows communication between HPC and Big Data systems by reducing the disturbance on HPC workloads while leveraging the built-in resilience of Big Data frameworks.

Processing tremendous volumes of data on dedicated big data technology is not as fast as processing the data on HPC infrastructure. This fact is recognized when comparing the efficiency of low-level programming models in HPC, which supports more parallelism, with big data technologies that are written with high-level programming languages. Many practical case studies and research have confirmed this fact. In [13], sentiment analysis on Twitter data was conducted for different dataset sizes using an MPI environment that showed better performance than using Apache Spark.

The enhancement of big data programming models can be achieved by integrating them with parallel programming models such as MPI. This approach can be seen in [4] that showed how to enable the Spark environment using the MPI libraries. Although this technique indicates remarkable speedups, it must use shared memory, and there are other overheads as a potential drawback. In [14], a scalable MapReduce framework, named Glasswing, is introduced. It is configured to use a mixture of coarse- and fine-grained parallelism to obtain high performance on multi-core

CPUs and GPUs. The performance of this framework is evaluated using five MapReduce applications with the indication that Glasswing outperforms Hadoop in terms of performance and resource utilization.

Data locality is a critical factor that affects both performance and energy consumption in HPC systems [15]. Many big data frameworks such as MapReduce and Spark can support this concept by sending the computation to where the data resides. In contrast, parallel programming models such as MPI lack this advantage. A novel approach by Yin et al. in [16], named DL-MPI, is proposed for MPI-based data-intensive applications to support data locality computation. It uses a data locality API that allows MPI-based programs to obtain data distribution information for compute nodes. Moreover, it proposes a probability scheduling algorithm for heterogeneous runtime environments that evaluate the unprocessed local data and the computing ability of each compute node.

In [17], a data distribution scheme is used by abstracting NUMA hardware peculiarities away from the programmer and delegating data distribution to a runtime system. Moreover, it uses task data dependence information, which is available with OpenMP 4.0RC2, as a guideline for scheduling OpenMP tasks to reduce data stall times.

Partitioning or decomposition is the first step for designing a parallel program by breaking down problems into small tasks. It includes two main types: domain or data decomposition and function decomposition [18]. These two types can be combined as mixed parallelism that employs an M-SPMD (multiple-single program multiple data) architecture, which includes both task parallelism (MPMD) and data parallelism (SPMD) [19].

The choice of decomposition type and parallelism paradigm is determined by resource availability. Furthermore, these resources may define the granularity level that the system can support [20]. There have been few empirical studies for performing data decomposition in the HPC field, [21] investigated this approach when designing parallel applications. Additionally, the state-of-practice was studied with probing tools used to perform this function. Moreover, a set of key requirements was derived for tools that support data decomposition and communication when parallelizing applications.

Based on this previous related work and to the best of our knowledge, there is no contribution yet that employs all the previous factors in terms of using a hybrid parallel programming model, decomposition technique and granularity approach to build a HPC-based Resource Management System for enhancing the performance of big data applications and optimizing HPC resource utilization. The contribution of this paper is to address this gap.

3. The Proposed Architecture

The proposed architecture will be built based on different techniques that will be integrated together to constitute a parallel HPC-based Resource Management System that enhances the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. In more details, the system will have the following techniques (Figure 2):

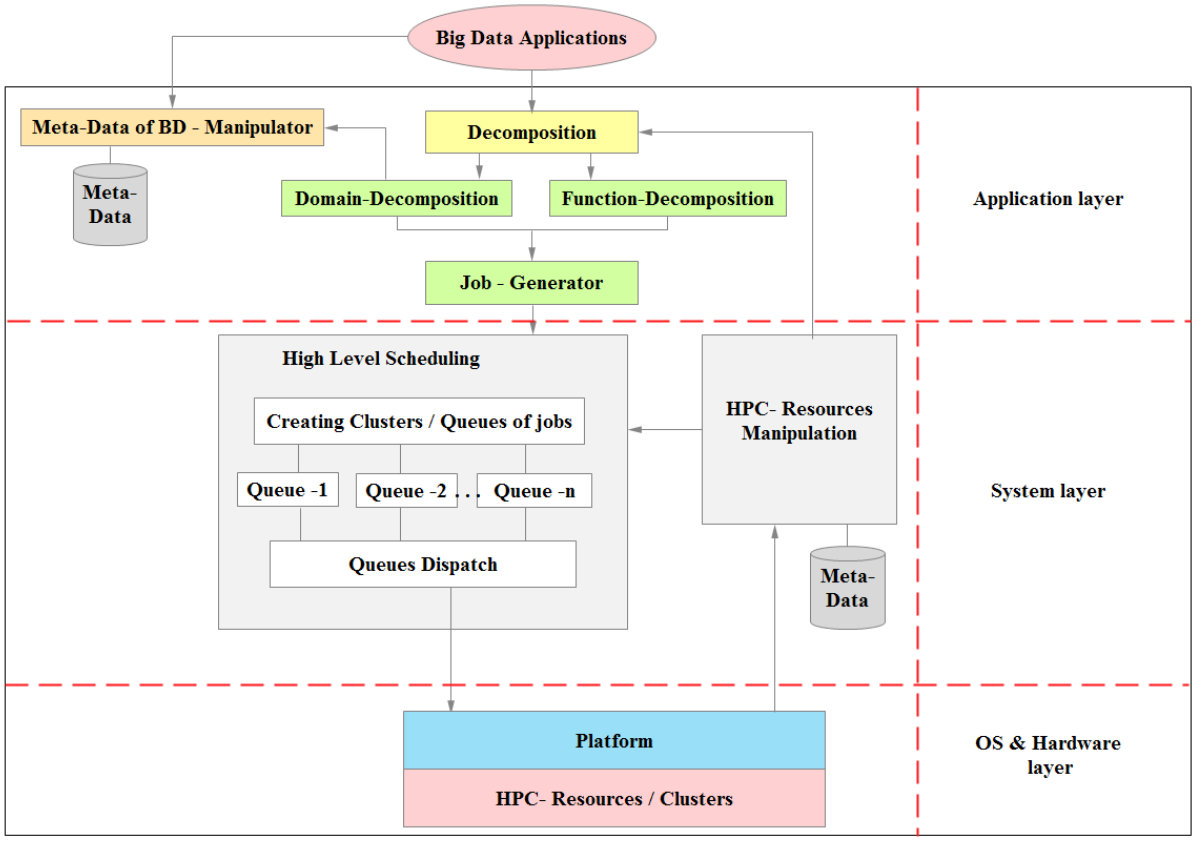


Figure 2: High-level architecture for the proposed system.

- **A new technique for HPC resources manipulation**

This dynamic technique will have some functionality related to HPC resources including:

  - Collecting a resource metadata to constitute a repository containing HPC resources with their capabilities and availabilities.
  - Tracking status of HPC resources.
  - Updating the metadata repository from time to time.

- **New decomposition techniques**

In this technique, we will provide a domain decomposition technique and/or a function decomposition technique for the given big data applications targeting both parallel computing architectures: Single Program Multiple Data (SPMD) and/or Multiple Instruction Multiple Data (MIMD). The availability of resources and system architecture can determine the decomposition paradigm and granularity options that efficiently support the resource management system.

- **A new high-level scheduling technique**

This technique will receive the decomposition results and creates clusters/queues of jobs during scheduling time based on the metadata about HPC resources/ clusters. It will consider data locality and load balancing while performing this task. Once clusters/queues of jobs are created, they will be dispatched to being executed by OS-platform on HPC resources/clusters.

4. Evaluation and comparative study

By comparing our architecture to other techniques, it is noticeable that the proposed architecture considers all the critical factors to achieve the performance and scalability attributes. Primarily, focusing on the topology awareness and building a metadata repository about the availabilities and capabilities of HPC resources can play a critical role to support the decomposition and high-level scheduling techniques. Such metadata can enhance the decision-making about choosing suitable granularity options and parallelism paradigm. Furthermore, the high-level scheduling technique can exploit the HPC resources positively by taking in account data locality and load balancing.

Instead of Integrating some big data and parallel programming models, this architecture constitutes an independent big data platform that employ hybrid parallel programming to support high parallelism for CPUs and GPUs accelerators.

The scalability can be seen from adding more dedicated clusters as needed. Adding more clusters will not affect the essence of each technique in particular, and resource management as a whole system integrating these techniques.

Different performance metrics have to be considered to implement the proposed architecture efficiently. Big data is the primary stream of this architecture, thus data building time is a significant metrics that may affect the performance. This time is required to construct a data structure used for computation and to perform the decomposition technique. Furthermore, employing parallel programming models such as MPI and OpenAcc can affect the computation time positively. From the part of HPC, hardware utilization metrics is also a cornerstone of the proposed architecture particularly for improving high-level scheduling technique

The novelty of this architecture can be arisen from having metadata about both big data applications and HPC resources, which leads to scheduling current jobs to the most suitable and available resources or cluster.

5. Conclusion

HPC has become an attractive environment for supporting scientific workflows and big data computing due to its performance capabilities. Unfortunately, big data applications usually do not fully exploit these capabilities afforded by HPC clusters, because such applications were written in high-level programming languages that do not encourage parallelism as parallel programming models. Another reason is that the architecture of big data platforms defers from the HPC architecture. A parallel HPC-based Resource Management System is proposed in this paper to enhance the performance of big data applications on HPC clusters without sacrificing the power consumption of HPC. For the future work, the High-level architecture for the proposed system will be developed and evaluated by running some big data applications. Moreover, some performance benchmarks will be provided to reflect the efficiency of our system.

References

[1] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[2] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[3] D. A. Reed and J. Dongarra, "Exascale computing and big data," *Commun. ACM*, vol. 58, no. 7, pp. 56–68, 2015.

[4] M. Anderson *et al.*, "Bridging the gap between HPC and big data frameworks," *Proc. VLDB Endow.*, vol. 10, no. 8, pp. 901–912, 2017.

[5] P. Xuan, J. Denton, P. K. Srimani, R. Ge, and F. Luo, "Big data analytics on traditional HPC infrastructure using two-level storage," *Proc. 2015 Int. Work. Data-Intensive Scalable Comput. Syst. - DISCS '15*, pp. 1–8, 2015.

[6] H. R. Asaadi, D. Khaldi, and B. Chapman, "A comparative survey of the HPC and big data paradigms: Analysis and experiments," *Proc. - IEEE Int. Conf. Clust. Comput. ICCC*, pp. 423–432, 2016.

[7] A. Reuther *et al.*, "Scalable system scheduling for HPC and big data," *J. Parallel Distrib. Comput.*, vol. 111, pp. 76–92, 2018.

[8] P. Jones, "NAS for Job Requirements Queuing / Scheduling Checklist Software," 2018.

[9] W. Saphir, L. A. Tanner, B. Traversat, W.~Saphier, L.A.~Tanner, and B.~Traversat, "Job Management Requirements for {NAS} Parallel Systems and Clusters," *IPPS Work. Job Sched. Strateg. Parallel Process.*, no. 949, pp. 319–336, 1995.

[10] Z. Wang and Y. Shen, "Job-Aware Scheduling for Big Data Processing," *Proc. - 2015 Int. Conf. Cloud Comput. Big Data, CCBD 2015*, pp. 177–180, 2016.

[11] C. Gómez-Martín, M. A. Vega-Rodríguez, and J. L. González-Sánchez, "Fattened backfilling: An improved strategy for job scheduling in parallel systems," *J. Parallel Distrib. Comput.*, vol. 97, pp. 69–77, 2016.

[12] M. Mercier, D. Glesser, Y. Georgiou, and O. Richard, "Big Data and HPC collocation : Using HPC idle resources for Big Data Analytics," pp. 347–352, 2017.

[13] D. S. Kumar and M. A. Rahman, "Performance evaluation of Apache Spark Vs MPI: A practical case study on twitter sentiment analysis," *J. Comput. Sci.*, vol. 13, no. 12, pp. 781–794, 2017.

[14] I. El-Helw, R. Hofman, and H. E. Bal, "Glasswing," *Int. Symp. High-Performance Parallel Distrib. Comput.*, pp. 295–298, 2014.

[15] D. Unat *et al.*, "Trends in Data Locality Abstractions for HPC Systems," *IEEE Trans. Parallel Distrib. Syst.*, pp. 1–1, 2017.

[16] J. Yin, A. Foran, and J. Wang, "DL-MPI: Enabling data locality computation

for MPI-based data-intensive applications," *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013*, pp. 506–511, 2013.

[17] A. Muddukrishna, P. A. Jonsson, and M. Brorsson, "Locality-aware task scheduling and data distribution for OpenMP programs on NUMA systems and manycore processors," *Sci. Program.*, vol. 2015, pp. 156–170, 2015.

[18] B. Ren, S. Krishnamoorthy, K. Agrawal, and M. Kulkarni, "Exploiting Vector and Multicore Parallelism for Recursive, Data- and Task-Parallel Programs," *Proc. 22nd ACM SIGPLAN Symp. Princ. Pract. Parallel Program. - PPoPP '17*, pp. 117–130, 2017.

[19] V. Boudet, F. Desprez, and F. Suter, "One-step algorithm for mixed data and task parallel scheduling without data replication," *Proc. - Int. Parallel Distrib. Process. Symp. IPDPS 2003*, no. October 2015, 2003.

[20] L. Silva and R. Buyya, "Parallel programming models and paradigms," *High Perform. Clust. Comput. Archit. ...*, pp. 4–27, 1999.

[21] A. Meade, D. K. Deeptimahanti, J. Buckley, and J. J. Collins, "An empirical study of data decomposition for software parallelization," *J. Syst. Softw.*, vol. 125, pp. 401–416, Mar. 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 4, 42-46 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Comparison of Electrogastrograms in Seated and Supine Positions Using Wayland Algorithm

Yasuyuki Matsuura[1,*], Hiroki Takada[2]

[1]*Department of Cross Cultural Studies, Gifu City Women's College, 501-0192, Japan*

[2]*Department of Human & Artificial Intelligence Systems, Graduate School of Engineering, University of Fukui, 910-8507, Japan*

ARTICLE INFO

Article history:
Received:18 December, 2018
Accepted:26 June, 2019
Online: 11 July, 2019

Keywords:
Electrogastrogram
Gastrointestinal motility
Postural Change
Wayland algorithm

ABSTRACT

In previous study, it is difficult to estimate the biological states such as postural ones and aetas in accordance with the previous analysis of electrogastrograms (EGGs), basically the frequency resolution of healthy individuals. In addition, few studies have compared gastric myoelectrical activity in supine and sitting posture. The aim of this study is to evaluate the severity of 3D sickness in daily life. The authors herein conduct nonlinear analysis of EGGs of the healthy young and give the feasibility to discriminate the postural differences. It is suggested that the EGGs acquired in the seated position have increased wavelength irregularity and attractor complexity than those acquired in the supine position.

1. Introduction

There are a myriad of activities in our sitting posture, including paper work, light work and others, and fatigue accumulates if this position is maintained for long periods. People usually sleep and respite in a supine position. Likewise, sick people and those affected by 3D sickness take a break in the supine posture. Moreover, humans hardly get motion sickness while maintaining a supine position [1]. In contrast, a seated position can lead to experience both motion sickness, especially in moving ships, cars, and trains, and visually induced motion sickness (VIMS) that occurs while viewing stereoscopic movies. Still, few studies have made a comparison between gastrointestinal motilities in supine and those in seated posture to evaluate the severity of the VIMS.

One of the most relevant theories for VIMS is the sensory conflict theory [2-4]. According to this theory, sensory mismatch between visual and vestibular signals would be induced in virtual space compared to the real world [5]. Moreover, close anatomical and electrophysiological relationships between the autonomic and vestibular nervous systems have been reported [6]. It has been known that there is interaction between vestibular and autonomic nervous systems in the physiological fields; the motion sickness might affect both systems because dizziness and hidropoiesis are induced by the reaction of the autonomic nervous system due to the excessive stimulus for the vestibular system. For instance, it has been reported that the level of histamine increased in the hypothalamus and brainstem in rats due to the rotation stimuli, which is possibly associated with vomiting during the motion sickness [7]. In addition, motion sickness seems to be induced by the contradiction among visual, vestibular, auditory, and proprioceptive inputs. These sensory informations can be compared with the previous experiences. The sensory informations are expected from the previous ones, however, they might be deviating from the previous experiences. According to the sensory conflict theory, the motion sickness would be occurred by this kind of the divergence.

Physiological and psychological measurements are used to evaluate the influence of VIMS on the body. The latter method to evaluate VIMS is the simulator sickness questionnaire [8]. On the other hand, physiological evaluation of VIMS has been conducted using parameters such as the high- and low- frequency components of electrogastrography, skin resistance, blood pressure, heart rate volubility, sweat rate, number of blinks, and respiratory rate [9–11].

A unique rhythm is associated with each variation of a physiological phenomenon. These include short repeated cycles, such as pulses, which are repeated in units of seconds, and long cycles, such as seasonal variations, which repeat in units of years. For instance, we can observe the circadian-dependent variability in the human blood pressure [12]. This is known as intra-day blood pressure fluctuation and can have an intra-day variation pattern characterized by high blood pressure in the daytime and decreased blood pressure at night [13]. Autonomous nerve activity

*Yasuyuki Matsuura, 7-1 Hitoichiba Kitamachi Gifu 501-0192 Japan, +81-58-296-3131 & matsuura@gifu-cwc.ac.jp

https://dx.doi.org/10.25046/aj040405

considerably affects this variability. Blood pressure increases or decreases depending on the variations in circulating blood volume and postural blood pressure reflex activity due to variations in posture. Heartrate and blood pressure intra-day variations are known to act as an evaluation index for autonomic nervous function. Additionally, the head-up tilt test is known as an autonomic nervous functional test associated with passive postural change [14]. Heart rate in healthy individuals increases, and the R -R interval estimated from the Electrocardiogram (ECG) decreases when using a tilt-table to switch from a supine to standing position [14]. Also, the same effects are reportedly observed when switching from a supine to seated position [15]. In these cases, the blood is accumulated in the lower extremities, however, vasoconstriction is automatically conducted to prevent from the hypostasis. The arterial baroreceptor is known as an sensory organ to detect the variations in blood pressure owing to the postural changes and other activities. This bio-signal reaches the brain stem vasomotor center (BSVC) through afferent pathways in the medium of vagus or glossopharyngeal nerve. The BSVC is activated with decline in the activity of the afferent pathways when the blood pressure decreases, and an effect is observed in excitability toward the sympathetic nerve; Noradrenaline is secreted from vascular wall sympathetic nerve endings, and adrenaline is, therefore, secreted from the adrenal medulla. Hence, the blood pressure can be maintained. As venous return decreases, the vagus nerve activity in the afferent pathway decreases through the low-pressure receptors in pulmonary veins and the atrium. As this information is received by the brain stem and reaches the hypothalamus, arginine vasopressin is secreted from the posterior pituitary, and an increased effect occurs in vasoconstriction and circulating plasma volume. Although cardiocirculatory research after postural changes is readily available, studies regarding the impact of posture changes on electrogastrograms are limited.

Percutaneous electrogastrography allows to noninvasively examine gastric myoelectrical activity. Human gastric myoelectrical activity cannot be measured by any other conventional method such as computed tomography, gastro-fiberscopes or magnetic resonance imaging. An electrogastrogram (EGG) is usually evaluated using Fourier analysis, running spectral analysis, and other spectrum analysis. However, the information retrieved by this analysis is limited, and EGGs are not as usual as electrocardiograms, electroencephalograms, and other biosignal measurements, thus compromising their development.

In 1921, Walter Alvarez [16] recorded the first human EGG. EGGs are recorded by placing Ag-AgCl electrodes on the surface of the epigastrium. EGG records the electrical signals which travel through the stomach muscles and control the muscle contractions. Human gastric slow wave or pacesetter potential activity generated by the interstitial cells of Cajal (ICCs) occurs at a frequency of 3 cycles per minute [17] [18], as shown in Figure 1. The ICCs are known as pacemaker cells that spontaneously depolarize and repolarize. The ICCs set the myoelectrical rhythmicity of the stomach and other areas of gastrointestinal tract. Recorded EGG signals that have noise with higher frequencies > 0.15 Hz must be filtered out [19]. These filters such as a low-pass filter, a band-pass filter eliminate most respiratory and cardiac rhythms in electromyogram.

Many previous studies on electrogastrography have been conducted but have been mostly related to the clinical field [20]. For instance, hormone, drugs, and motion sickness effects on EGG have been evaluated. Likewise, EGG has been used to study the

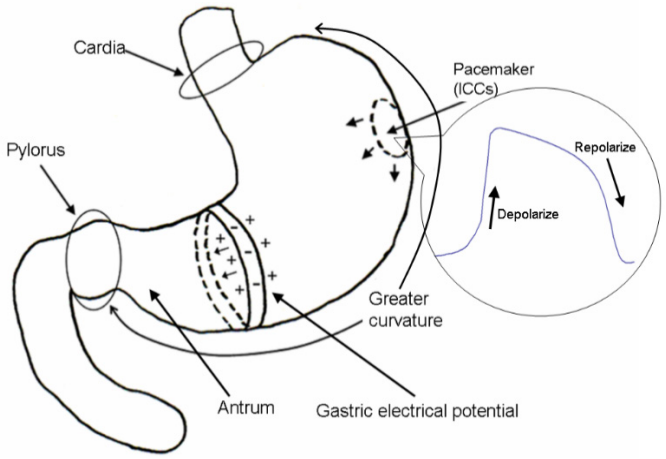


Figure 1: Traveling slow waves. The ICCs set the myoelectrical rhythmicity of the stomach and other areas of gastrointestinal tract.

effects of warm compresses on both gastrointestinal activity [21] and the epigastric region for relieving constipation [22], and to characterize the intestinal activity in patients with chronic constipation [23]. However, studies on clinical applications mainly consider patients, and hence scarce research is available on the EGG dynamics of healthy subjects.

More specifically, few studies have been focused on gastrointestinal motility in seated and supine positions to evaluate the severity of motion sickness. As it has been reported that three-dimensional motion sickness is strongly elicited when viewing stereoscopic movies for prolonged periods, it is important to assess the safety while viewing such movies. Therefore, in this study the authors compared EGGs in seated and supine positions using the Wayland algorithm to evaluate motion sickness induced by stereoscopic movies.

2. Material and Method

The participants were 5 healthy females aged 19 to 24 years. The Ethical Committee of Graduate School of Natural Sciences, Nagoya City University approved the experiment. Prior to the experiment, each participant signed their informed consents.

The design of experiment is Figure 2. This experiments was conducted, considering the order effect in the protocol. In Addition, postural change on experiment B is passive position change to remove the effect of muscle sympathetic nerve activity.

The experimental room and chair condition is Table 1. The experimental condition is Table 2. Participants finished their meals 120 mins beforehand. Therefore, the experiments were not affected by the presence of food.

Electrodes of EGG pasted on abdominal surface using 2 disposable Ag-AgCl electrodes as shown in Figure 3. Their abdominal surface of the pasting points was cleaned out with cotton moistened with ethanol and the skin abrasion (SkinPure, Nihon Kohden, Japan).

Using A/D converter, the analog signals of their EGGs were A/D-converted to digital signals at 1 kHz. A low-pass filter of 0.15 Hz eliminated the activity in respiratory, cardiac rhythms, electromyogram, and the noise of electronic devices. In addition, resampling was performed every 1 Hz to remove the noise in the abovementioned measurement.

Experiment A (Seated)

Seated Posture (Rest) 30 min	Seated Posture 90 min

t_0

Experiment B (Supine)

Seated Posture (Rest) 30 min	Supine Posture 90 min

t_0

Figure 2: Experiments A and B

Table 1: Experimental room and chair condition

Items	Conditions
Room Condition	Without windows
Sound-insulated	40dB
Room temperature	20–24°C
Air current	below 0.1 m/s
Chair size (Seating surface)	55cm (width) × 55cm (Length)
Chair size (Seating back)	55cm (width) × 58cm (Length)
Chair angle between the seating surface and the back	100 degrees

Table 2 Experimental condition and machines

Items	Condition/Machine
Measurement Start time	14:00 - 15:00
EGGs Sampling Rate	1KHz
Bio-amplifier	MT11; NEC Medical, Japan
Recorder	PC216Ax; Sony Precision Technology, Japan
A/D converter	AD16-16U (PCI) EH; CONTEC, Japan
Electrodes	Vitrode Bs, Nihon Kohden, Japan

Every 300-second, we analyzed time series data of the EGG for a 1024-second time window to divide the data (Figure 4). In this study, the authors calculated the translation error (E_{trans}) using Wayland algorithm from EGG time series data in 1 to 10 dimensional phase space [24-26].

3. Results

EGGs in seated (experiment A) and supine (experiment B) positions of healthy women over 5 min after 70 min from the beginning of measurements are shown in Figures 5(a) and (b), respectively. The EGG in seated position (Figure 5(a)) shows a large amplitude and unstable fluctuations. In contrast, the EGG in supine position (Figure 5(b)) shows a regular pattern. Figures 6(a) and (b) show the two-dimensional attractors ($\tau = 3$) generated from the EGGs in seated (Figure 5(a)) and supine (Figure 5(b)) positions, respectively.

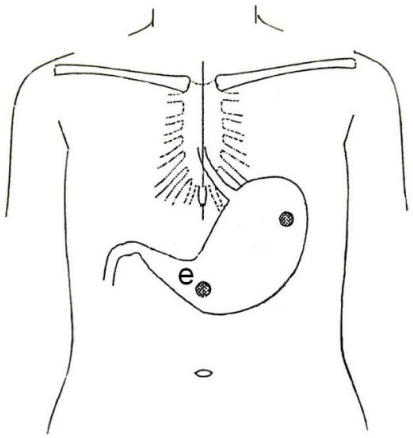


Figure 3: Electrodes position pasted on the surfece of body.

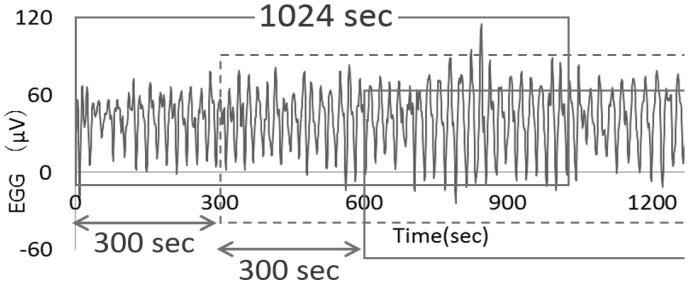


Figure 4: Moving segment of EGG-data.

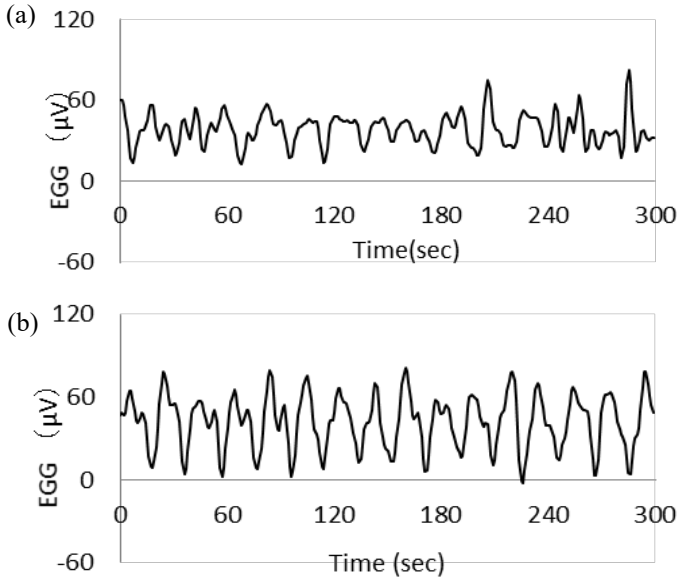


Figure 5: Typical examples of EGG in (a) experiment A and in (b) experiment B.

The EGG translation errors over a ten-dimensional embedding space in the seated position (excluding control) ranged from 0.27 to 0.58, with average of 0.46 and standard deviation of 0.07. The similar EGG translation errors in the supine position ranged from 0.23 to 0.52, with average of 0.40 and standard deviation of 0.08. The authors compared the translation errors estimated from the EGGs of experiment A with those of experiment B. Variations of z-scores of translation error (average ± standard error) were

estimated from the EGGs of the individuals in the ten-dimensional embedding space (Figure 7). The *z*-scores in the seated position (excluding control) ranged from –2.48 to 2.45, with average of 0.07 and standard error of 0.42, and those in the supine position ranged from –1.63 to 3.18, with average of 0.04 and standard error of 0.42.

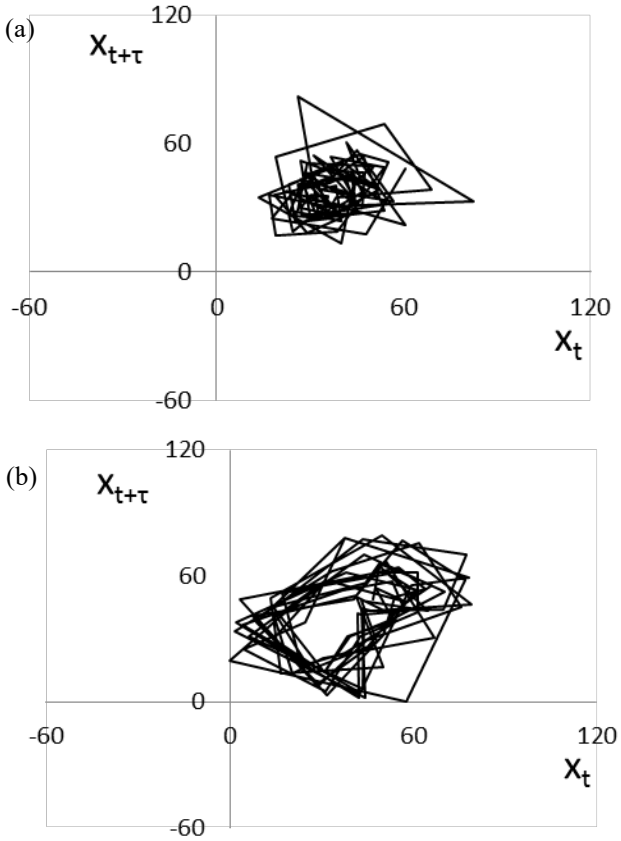


Figure 6: Attractors of EGG in (a) experiment A (Figure 5(a)) and in (b) experiment B (Figure 5(b)).

Then, the values of the translation errors estimated from the EGGs of experiments A and B were compared using the Wilcoxon signed-rank test. Significant differences were observed 50 and 55 min after the end of the resting period with $p < 0.05$. The authors confirmed that translation errors in the ten-dimensional embedding space are sufficiently stable compared to those in embedding spaces with lower dimension. Figure 8 shows variations of the difference between the translation error estimated from the EGG at time t and that at onset time t_0, except for the resting period in the ten-dimensional embedding space. The variations were calculated as *DTE* in the seated position (excluding control) ranged from –0.19 to 0.13, with average of –0.001 and standard error of 0.035, and that in the supine position ranged from –0.13 to 0.086, with average of –0.011 and standard error of 0.021.

4. Discussion

To estimate the EGG dynamics, the authors analyzed EGG signals using the Wayland algorithm. The regularity and periodicity of the EGGs was considered the cause of the abovementioned results. As shown in Figures 7 and 8, significant differences in translation errors estimated from EGGs in seated and supine positions were determined after 50 and 55 min from the resting period.

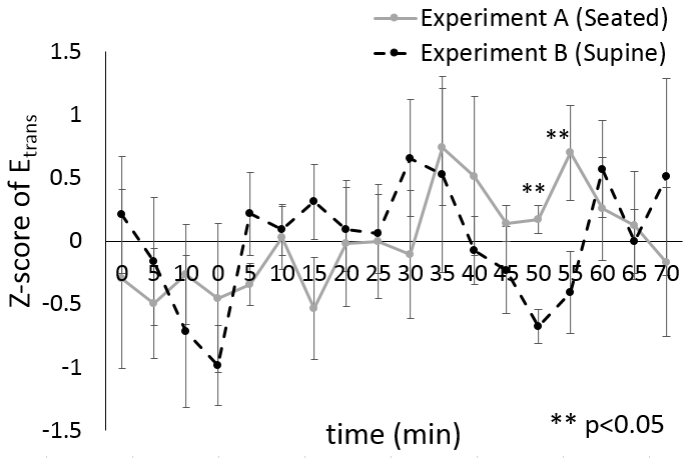


Figure 7: Z-scores of translation error (average ± standard error).

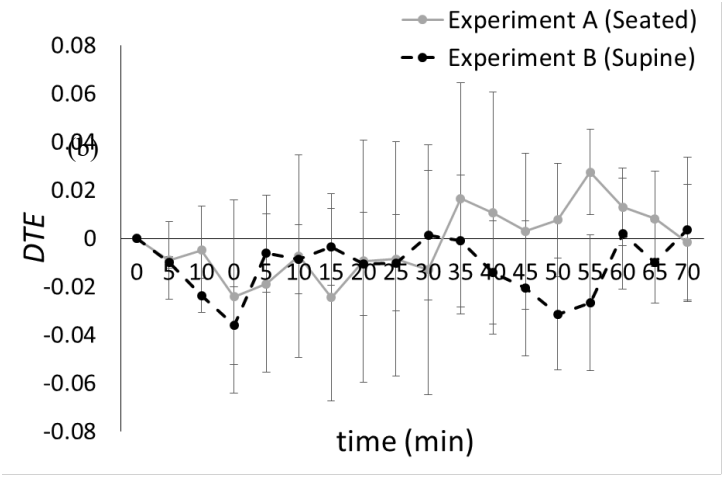


Figure 8: Difference translation errors (average ± standard error).

$$DTE(t) = E_{trans}(t) - E_{trans}(t_0).$$

Some studies have reported about the effect of postural change on the cardiocirculatory system after subjects changes from a supine to seated/standing postures, however, few does not have studied the effect of postural change on the autonomic nervous system, especially on the EGGs recorded over a prolonged period. Measuring bowel sounds, Kobayashi et al. tried to investigate the effect of the postural changes on the intestinal peristalsis [27]. Healthy middle-aged women voluntary participated in this study. Their bowel sounds were recorded during the right lateral recumbent following the supine for an hour. The results showed that intestinal peristalsis was significantly rose 30-75 min after the postural change whereas these results are questionable owing to the considerable differences prevalent among the participants [27]. In the other previous work, it has been reported that intestinal peristalsis would be temporarily suppressed if the intra-abdominal organ were affected by sudden postural changes [28]. Recording the bowel sounds 30 min after the postural change to the supine is appropriate for stationary measurements [29]. These results suggest that the stomach or intestines become active by means of parasympathetic nerve activity acceleration after a determined amount of time has passed since resuming supine position.

In general, EGGs are recorded in either a seated or supine position [30], and motion sickness is more often induced in the seated than in the supine position. In addition, the effect of postural

change on the autonomic nervous system, especially in the digestive system, has been investigated in some studies, although with focus on the cardiovascular system [31][32]. In this study, the authors conducted a basic experiment to investigate the effect of postural change on the autonomic nervous system with focus on the digestive system.

5. Conclusions

Motion sickness and diseases are usually settled in a seated position. In fact, humans can hardly get motion sickness in a supine position. This study aimed to compare EGGs acquired in seated and supine positions using the Wayland algorithm to evaluate motion sickness induced by stereoscopic movies. The authors verified that the supine position retrieves more stable EGG signals in the supine than in the seated position, thus confirming the benefits of the supine position to relieve motion sickness. There are various cycles that occur in biorhythms, and many of the variations are affected by regional differentiation of the sympathetic system. Biorhythms are also impacted by variations in posture as well as the physical condition of a person and their lifestyle. Integrative and multifaceted evaluation of impact from biorhythm and posture variations is therefore useful for evaluating autonomic nervous function and estimating biological condition. The lack of research on gastric motility and intestinal peristalsis due to variation in posture warrants an ongoing necessity for accumulation of foundational research. In upcoming studies, the authors will measure heart rate, blood pressure and multichannel EGG signals in seated and supine positions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was supported in part by the Kayamori Foundation of Informational Science Advancement, the Japan Society for the Promotion of Science, and Grant-in-Aid for Scientific Research (C) (Grant numbers 17K00715 and 18K11417).

References

[1] G.W. Manning, G.W. Stewart, “Effect of body position on incidence of motion sickness,” J Appl Physiol. 1(9), 619-628, 1949.
[2] J.T. Reason, J.J. Brand, Motion Sickness, 3rd ed, San Diego: Academic Press Inc, 1975.
[3] J. A. Irwin, "The pathology of sea-sickness," The Lancet. 118(3039), 907-909, 1878.
[4] B. Keshavarz, H. Hecht, B. D. Lawson., "Visually induced motion sickness: Causes, characteristics, and countermeasures," in Handbook of Virtual Environments, 2nd Edition, K. S. Hale, K. M. Stanney, Ed, FL, CRC Press, 2015, 647-698.
[5] N.H. Barmack, “Central Vestibular System: Vestibular nuclei and posterior cerebellum,” Brain Res Bull 60(5-6), 511-541, 2003.
[6] C.D. Balaban, J.D. Porter, “Neuroanatomical substrates for vestibuloautonomic interactions,” J Vestib Res. 8, 7-16, 1998.
[7] N. Takeda, M. Morita, T. Kubo, A. Yamatodani, T. Watanabe, H. Wada, T. Matsunaga, “Histaminergic mechanism of motion sickness neurochemical and neuropharmacological studies in rats,” Acta Otolaryngol. 101(5-6), 416-421, 1986.
[8] R.S. Kennedy, N.E. Lane, K.S. Berbaum, M.G. Lilienthal, “Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness,” Int J Aviat Psychol. 3(3), 203-220, 1993.
[9] S.R. Holmes, M.J. Griffin, “Correlation Between Heart Rate and the Severity of Motion Sickness Caused by Optokinetic Stimulation,” J Psychophysiol, 15, 35-42, 2001.
[10] N. Himi, T. Koga, E. Nakamura, M. Kobashi, M. Yamane, K. Tsujioka, “Differences in autonomic responses between subjects with and without nausea while watching an irregularly oscillating video,” Auton Neurosci. 116(1-2), 46-53, 2004.
[11] Y. Yokota, M. Aoki, K. Mizuta, Y. Ito, N. Isu, “Motion sickness susceptibility associated with visually induced postural instability and cardiac autonomic responses in healthy subjects,” Acta Otolaryngol. 125(3), 280-285, 2005.
[12] R.C. Hermida, D.E. Ayala, J.R. Fernandex, L.M. Ruilope, J.E. Lopez, “Modeling the circadian variability of ambulatorily monitored blood pressure by multiple-component analysis”, Chronol Int., 19, 461-481, 2002.
[13] A.M. Birkenhanger, A.H. van den Meiracker, “Causes and consequences of a non-dipping blood pressure profile”, Netherland J Med., 65, 127-131, 2007.
[14] S. Iwase, T. Mano, M. Saito, “Effects of graded head-up tilting on muscle sympathetic activities in man”, Physiologist, 20, S62-S63, 1987.
[15] H. Noro, I. Watanabe, N. Yabunaka, Y. Ohtsuka, Y. Agishi, “Quantitative Evaluation of Changes in pulse Rate and Blood Pressure by the Supine - to - Sitting active Postural Change Test”, The Autonomic nervous system 32(2), 111-118, 1995.
[16] W.C. Alvarez, “The electrogastrogram and what it shows,” JAMA. 78(15), 1116-1118, 1922.
[17] D. Couturier, C. Roze, J. Paolaggi, C. Debray, “Electrical activity of the normal human stomach: A comparative study of recordings obtained from the serosal and mucosal sides,” Dig. Dis. Sci. 17(11) 969-976, 1972.
[18] E.J. Van Der Schee, A.J.P.M. Smout and J.L. Grashuis, “Application of running spectrum analysis to electrogastrographic signals recorded from dog and man,” in Motility of the digestive tract, M. Wienbeck, Ed, New York: Raven Press, 1982, 241-250.
[19] J.Z. Chen, R.W. McCallum, Electrogastrography: Principles and Applications, New York: Raven Press, 1994.
[20] M. Nagai, M. Wada, Y. Kobayashi, S. Togawa, “Effects of lumbar skin warming on gastric motility and blood pressure in humans,” Jpn. J. of Physiol. 53(1), 45-51, 2003.
[21] N. Kawachi, S. Iwase, H. Takada, D. Michigami, Y. Watanabe, N. Mae, “Effect of wet hot packs applied to the epigastrium on electrogastrogram in constipated young women,” Autonomic Nervous System 39(5), 433-437, 2002.
[22] Y. Matsuura, S. Iwase, H. Takada, Y. Watanabe, E. Miyashita, “Effect of three days of consecutive hot wet pack application to the epigastrium on electrogastrography in constipated young women,” Autonomic Nervous System. 40(4), 406-411, 2003.
[23] S.R. Cajal, Histologie du systeme nerveux de l'homme et des vertebras, Vol. 2. Paris :Maloine, 1909, p. 942. (in French); S.R. Cajal, N. Swanson, L.W. Swanson, Histology of the nervous system of man and vertebrates, New York: Oxford University Press, 1995. (reprinted in English)
[24] R. Wayland, “Recognizing Determinism in a Time Series,” Phys. Rev. Lett. 70-5, 580-582, 1993.
[25] H. Takada, T. Morimoto, H. Tsunashima, T. Yamazaki, H. Hoshina, M. Miyao, “Applications of Double-Wayland Algorithm to Detect Anomalous Signals,” Forma, 21(2), 159–167, 2006.
[26] Y. Matsuura, H. Takada, K. Yokoyama, K. Shimada, “Proposal for a New Diagram to Evaluate the Form of the Attractor Reconstructed from Electrogastrography”, Forma, 23, 25-30, 2008.
[27] T. Kobayashi, T. Yazaki, Y. Hujihara, H. Ozawa, J. Nakahashi, “Analysis of Change in Bowel Sound and Movement by the Posture Change: Using an Actigraph Micromini Type Sound Sensor”, Bulletin of Faculty of Nursing, Yamanashi Prefectural University 13, 37-46, 2011.
[28] K. Fukai, N. Sakamoto, N. Tanaka, “Effect of Oral Water Intake, Exercise and Hot Compresses on Bowel Sounds in Healthy Women”, Kawasaki medical welfare journal 6(1), 99-106, 1996.
[29] N. Sakamoto, K. Fukai, N. Tanaka, “The Relationship Between Bowel Sounds and Constipation Assessment in Healthy Women”, Kawasaki medical welfare journal 6(2), 341-346, 1996.
[30] K. L. Koch, R. M. Stern, Handbook of Electrogastrography, New York: Oxford University Press, 2004.
[31] K. Nishijo, “The Relation between the Penpheral Clrculatory Behaviour and the Changes of the Body Posrtron”, Bulletin of defectology, 2, 33-45, 1978.
[32] S. Mukai, J. Hayano, “Heart rate and blood pressure variabilities during graded head-up tilt”, J. Appl. Physiol., 78(1), 212-216, 1995.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 1, 45-49 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Morphological and Optoelectrical Characterization of Silicon Nanostructures for Photovoltaic Applications

Babacar Dieng[*,1], Moussa Toure[2], Modou Beye[1], Diouma Kobor[2], Amadou Seidou Maiga[1]

[1]*Laboratory of Electronic, Informatics, Telecommunication and Renewable Energy (LEITER), Gaston Berger University of Saint-Louis, Senegal*

[2]*Laboratory of Chemistry and Physics of Materials (LCPM), Assane Seck University of Ziguinchor, Senegal*

ARTICLEINFO

Article history:
Received: 14 August, 2018
Accepted: 23 December, 2018
Online: 21 January, 2019

Keywords:
Silicon nanostructures
Metal assisted chemical etching
Optoelectrical characterization

ABSTRACT

Metal (silver)-Assisted Chemical Etching (MACE) method is used to fabricate silicon nanostructures like silicon nanowires (SiNWs) and silicon nanocones (SiNCs). The morphological characterization of fabricated SiNWs has shown that 5 minutes is the optimal time of silver deposition on silicon substrate. Silicon nanocones (SiNCs) were also fabricated by etching vertical SiNWs with a $AgNO_3/HF/H_2O_2$ solution. The optical and electrical properties of SiNWs and SiNCs are analyzed and compared with those of the bulk silicon. The fabricated SiNWs (SiNCs) reduce the surface reflectance and the sheet resistance down to 6% (3%) and 9.143 Ω/sq (6.997 Ω/sq) respectively.

1 Introduction

The energy demand worldwide is largely greater than the energy supply. Thus, the conventional energy sources (oil, coal and nuclear) become expensive. In the other way, these conventional energies have negative impact on the environment and the human health. The explosions of the Chernobyl and Fukushima nuclear power stations cause damage on the human health in many countries [1][2]. Consequently, the development of alternative energy sources is urgently needed. Solar energy, as a clean, abundant and renewable source, is one of the most promising alternatives to conventional energies.

The incident light is converted into electric current by photovoltaic (PV) cells. The PV cells are fabricated by using semiconductor materials. Among the semiconductor used in the PV cell fabrication, the bulk silicon (mono or multi-crystalline) is mostly dominated. The best conversion efficiencies obtained with monocrystalline (multi-crystalline) silicon solar cells and modules are around 26.7% (21.9%) and 24.4% (19.9%) respectively [3]. However, the crystalline silicon represent about 40% of the production cost of bulk silicon modules [4].

Further cost reduction and efficiency enhancement, through reduction of material usage, simplification of device fabrication as well as optimization of device structure, are required. A way to take up this challenge is to use silicon nanostructures, which theoretically should be both more efficient and less expensive [5].

The common methods to fabricate silicon nanostructures (SiNWs, SiNCs) are dry etching and lithography methods [3][6]. However, these methods are complex, expensive, and unsuitable for mass production. One of the alternative methods is to use Metal Assisted Chemical Etching (MACE) which is based on the strong catalytic activity of metals in HF/H_2O_2 aqueous solution [7]. This low cost method is very easy to implement and it allows to realize various silicon nanostructures with excellent optoelectronic properties without changing the purity of silicon.

Dimova Malinovska et *al.* [8] are the first to etch the silicon by using MACE method. They demonstrated that porous silicon can be obtained by using aluminum metal and $HF/HNO_3/H_2O$ solution. Since 2002, the MACE method is being improved and a variety of nanostructures are fabricated. SiNWs were firstly fabricated by Peng et al. They used the mixture of $HF/AgNO_3$ [9][10]. Vertical SiNWs were fabricated by combining MACE and lithography methods [11]. However, the cost of these nanowires is relatively high because of the lithography technique used to deposit the PS sphere masks. Moreover, Au metal was used as catalyst. To realize best silicon nanostructures with-

[*]Babacar Dieng, Gaston Berger University of Saint-Louis, Senegal, Email: dieng.babacar@ugb.edu.sn

https://dx.doi.org/10.25046/aj040106

out using any mask patterns and expensive metals is required in the solar cell manufacturing.

The objective of this work is to fabricate silicon nanostructures (nanowires and nanocones) by MACE method in ambient temperature, without using any mask patterns, and to study their structural and opto-electronic properties.

2 Materials and method

The silicon nanostructures were fabricated by using (100) p-type monocrystalline silicon substrate. These samples have a resistivity and a thickness in the ranges of 1-5Ω/cm and 600 - 650 μm, respectively.

The different steps of the SiNWs fabrication by MACE are summarized in the figure 1. First, the samples were cleaned by using, respectively, acetone, ethanol and deionized water in the ultrasonic bath for 15 minutes. After this cleaning, the samples were immersed in a piranha solution (H_2SO_4/H_2O_2) for 10 minutes to eliminate any organic trace. The last step of the cleaning process was to use a diluted HF solution to remove the native silicon oxide SiO_2 and to rinse the samples with deionized water. After the cleaning steps, the samples were immersed in the $HF/AgNO_3$ (4.8M/0.02M) solution for a chosen time in [0.5, 2.5, 5, 10, 15, 30, 60, 120 minutes]. This solution allows to deposit silver (Ag) nanoparticles on the silicon surface. After depositing Ag nanoparticles, samples were directly immersed in HF/H_2O_2 (4.8M/1.176.10^{-3}M) mixture. This solution is responsible for the etching silicon. After the etching process, all Ag waste integrated into the silicon substrate must be removed. For this, a diluted HNO_3 solution was used.

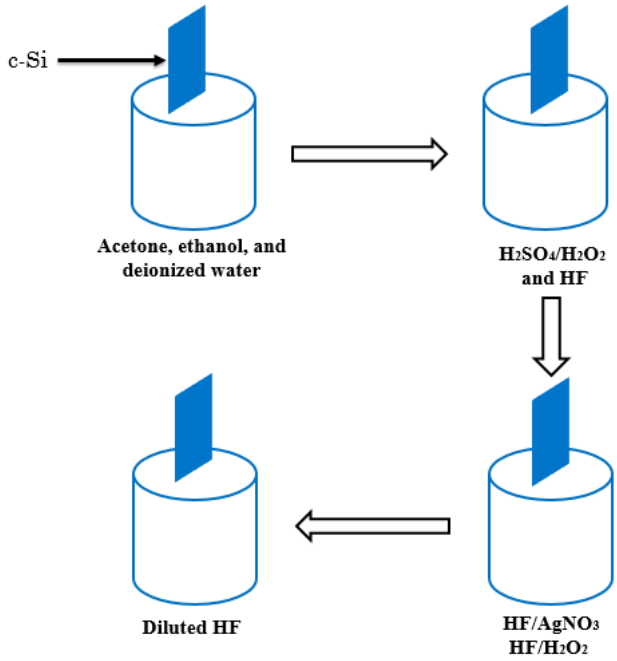


Figure 1: Different steps in the fabrication of the silicon nanowires by MACE

The conical form of silicon nanostructures can be fabricated by MACE via obtained silicon nanowires (SiNWs). For this, SiNWs are introduced into $AgNO_3/HF/H_2O_2$ solution. This solution have double functions. First, it deposit Ag nanoparticles on the top of SiNWs. And second, it etch these SiNWs at the top to form truncated SiNWs like silicon nanocones (SiNCs).
It is interesting then to understand how these nanostructures occur on the surface of silicon substrate.

3 Mechanism of the formation of silicon nanostructures

The growth mechanism of SiNWs is described by Peng et *al.*[12]. The Ag^+ ions given by $HF/AgNO_3$ seize the valence band electrons of silicon. Thus, these ions become now Ag nucleation. Ag nuclei gradually grow into Ag nanoparticles. The reduction of H_2O_2 oxydant is done at Ag nanoparticles; and the holes are consumed by the silicon oxidation. The produced silicon oxyde is quickly dissolved by HF. Ag nanoparticles travel into the silicon wafer as the silica layer is dissolved, thus the depth of Ag nanoparticles in the pit increase gradually with the increase of the reaction time.

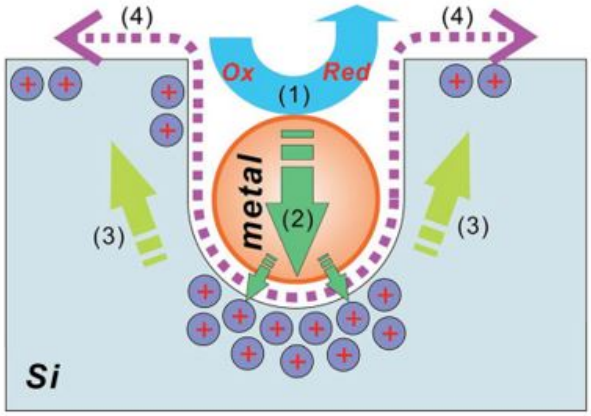


Figure 2: Formation mechanism of silicon nanowires fabrication: (1) reduction of oxidant H_2O_2, (2) injection of generated holes onto the silicon substrate, (3) migration of holes inside the silicon surfaces, and (4) dissolution of silicon oxide by HF [13]

The Ag nanoparticle side in contact with the etching solution acts as a cathode and serves to reduce H_2O_2 (see the equation 1). The other side of Ag particles which is in contact with the silicon substrate functions as an anode that serves to oxidize silicon which generates H^+ and electrons (see equations 2 to 4). A considerable difference of potentials between the cathode and anode sites is then created and therefore a local oxidation current flows from the cathode site to the anode one.

$$H_2O_2 + 2H^+ + 2e^- \longrightarrow 2H_2O \quad (1)$$

$$Si + 2H_2O \longrightarrow SiO_2 + 4H^+ + 4e^- \quad (2)$$

$$SiO_2 + 6HF \longrightarrow H_2SiF_6 + 2H_2O \quad (3)$$

$$Si + 6HF \longrightarrow H_2SiF_6 + 4H^+ + 4e^- \quad (4)$$

The overall equation is given by:

$$Si + 2H_2O_2 + 6HF \longrightarrow H_2SiF_6 + 4H_2O \quad (5)$$

The same reactions can explain the formation of silicon nanocones by etching the SiNW tops. Indeed, when the sample is put in $AgNO_3/HF/H_2O_2$ solution, Ag nanoparticles are deposited on the SiNWs. Thus, H_2O_2 oxydises the SiNW tops and silicon oxyde is reduced by HF. These oxydation and dissolution will occur at the same time during the etching process. Then, silicon nanocones will be formed on silicon substrate by truncating the edge top of SiNWs, according to the study of Shimizu et *al.*[12].

A MERLIN Scanning Electron Microscope (SEM) with Zeiss FEG was used for structural characterization of the elaborated silicon nanowires. The spectral reflectance of fabricated silicon nanostructures is measured by Hitachi UV-VIS-NIR 4001 spectrophotometer equipped with an integrating sphere. Their electrical properties are studied by measuring the sheet resistance with 4-point probes method.

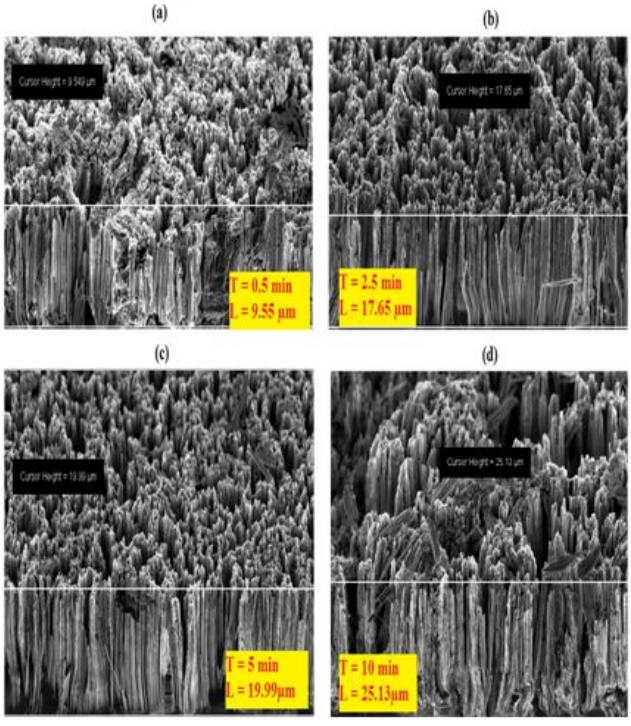


Figure 3: SEM images of silicon nanowires fabricated by the MACE method

4 Results and discussions

Figures 3 and 4 show the SEM characterization results of the fabricated SiNWs. The SiNW height (L) depends on the deposition time of silver nanoparticles (T). The SEM characterization results show that L increases with T. Therefore continuous film of Ag nanoparticles can be obtained by increasing T. This continuous film with a few holes permits to obtain well-defined SiNWs [13]. However, a spongy structure is shown when T is very high (more than 30 minutes). It can be explained by the break of long SiNWs. Moreover, a structure with too much spaced SiNWs, is obtained for too long deposition time of silver nanoparticles. This spacing of nanowires as function of the etching time could be due to the phenomenon of the coalescence of silver nanoparticles which leads to the etching of the silicon under the silver macroparticles.

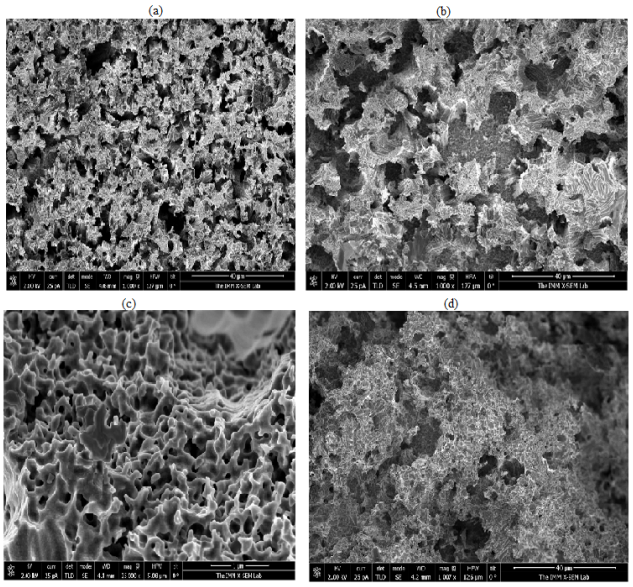


Figure 4: SEM images of silicon nanowires fabricated by MACE method for (a) 15 minutes, (b) 30 minutes, (c) 60 minutes and (d) 120 minutes.

The reflectance of the fabricated SINWs was measured and the results are shown in figure 5. In this figure, the reflectance of bulk silicon is used as the reference. All SiNWs give very low reflectance in the useful wavelength range. The decrease of reflectance is possible by increasing the SiNWs height. This occurs because very long SiNWs lead to multiple rebounds of the light inside the structure and to the increase of the probability of light absorption. However, the SiNWs must be vertical to the substrate as well found in Figure 3c. Thus, the 5-minutes processed SiNWs give the best antireflective properties. They allow to reduce the surface reflectance of silicon bellow 6% in the 400-1100 nm wavelength range.

In order to improve the silicon antireflective properties, nanocone arrays were formed on silicon substrate. The 5-minutes processed SiNWs were used to fabricate silicon nanocone arrays because they give the best optical and structural results (figures 3c and 5). The optical characterization results of the fabricated nanocone arrays are shown in the figure 6. According to these results, all the nanocones arrays give the lowest reflectance compared to the SiNWs one. The reflectance around 3% is obtained with the 5-minutes fabricated nanocone arrays.

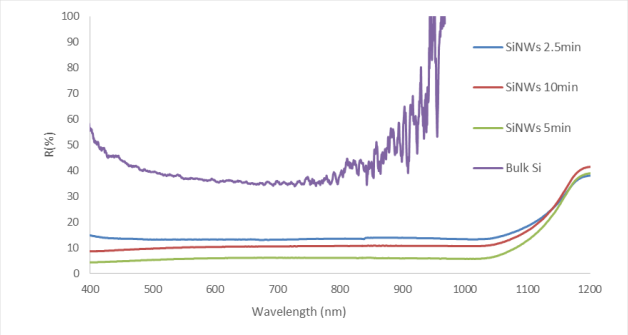


Figure 5: Reflectance of bulk silicon and silicon nanowires (SiNWs).

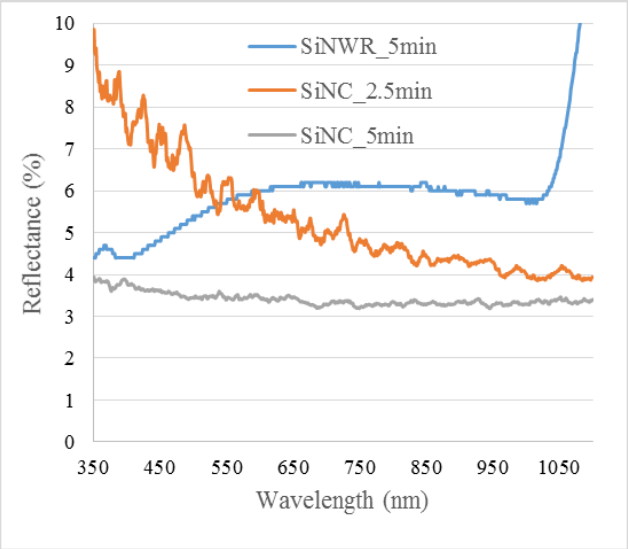


Figure 6: Reflectance of fabricated silicon nanocones (SiNC) compared with SiNW one.

The large index mismath between air and silicon substrate can be reduced by using ergonomic tapered nanostructures which grade the refractive index from silicon to the air. This grading refractive index is observed for all nanostructure arrays, specially nanocone arrays. They allow to obtain a volume fraction variation between the top and the bottom of nanocones. Furthermore, the nanostructures diffract incoming light and increase the photon path length within the absorber layer.

A comparative study of the electrical properties was also carried out between the bulk monocrystalline silicon, and silicon nanostructures (SiNWs and SiNCs). The surface resistivity called also sheet resistance (R_{sheet}) permits to evaluate the electrical conductivity of a nanostructured surface. It can be defined as the ratio of the voltage drop per unit length to the surface current per width. The common instrument to measure the sheet resistance is the four-point probes method. It consists to apply a current (I) on the outer two probes and to measure the resultant voltage drop (ΔV) between the inner two drop. Thus, the sheet resistance is calculated by the following equation [14]:

$$R_{sheet} = \frac{\pi}{\ln(2)} * \frac{\Delta V}{I} \quad (6)$$

Ohm-per-square (Ω/sq) is the measurement unit of the sheet resistance. The resistivity of sample can be defined as the product of their sheet resistance and thickness. According to the results shown in the figures 7 to 9, the sheet resistance of silicon is reduced by nanostructuring the silicon substrate surface. The sheet resistance of 9.143 and 6.997Ω/sq have been obtained for silicon nanowires and silicon nanocones respectively.

A material with a low sheet resistance is better to transfer the electrical charge. Furthermore, the resistivity and conductivity can be calculated if the sheet resistance and material thickness are known. This allows for the materials to be electrically characterized, by measuring their sheet resistance. Therefore, the electric conductivity of SiNWs and SiNCs (with low sheet resistances) are far better than that of bulk silicon.

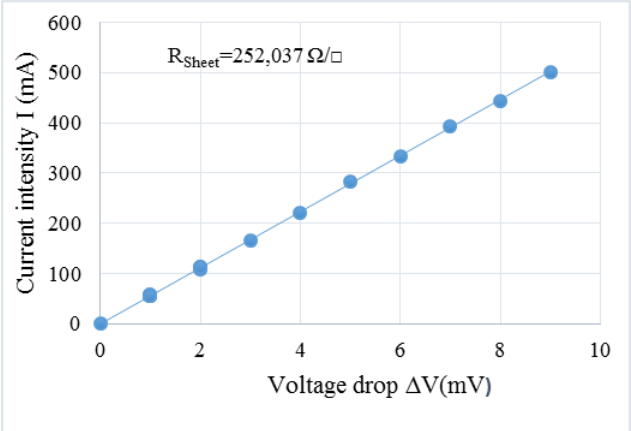


Figure 7: Variation of I(mA) as function of ΔV(mV) of bulk monocrystalline silicon.

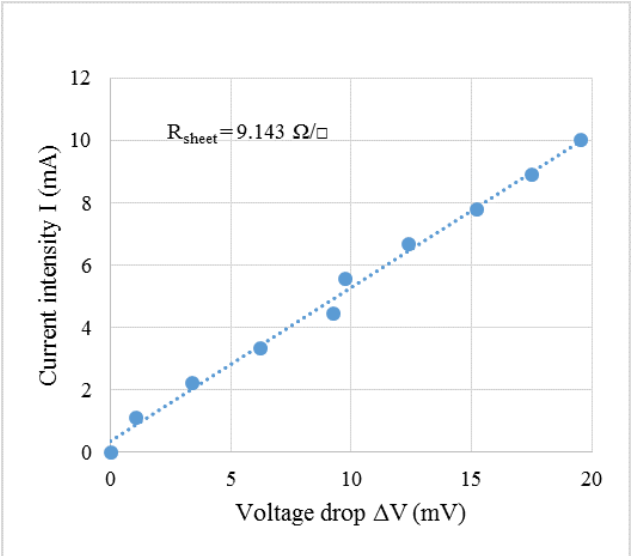


Figure 8: Variation of I(mA) as function of ΔV(mV) of 5-minutes fabricated silicon nanowires.

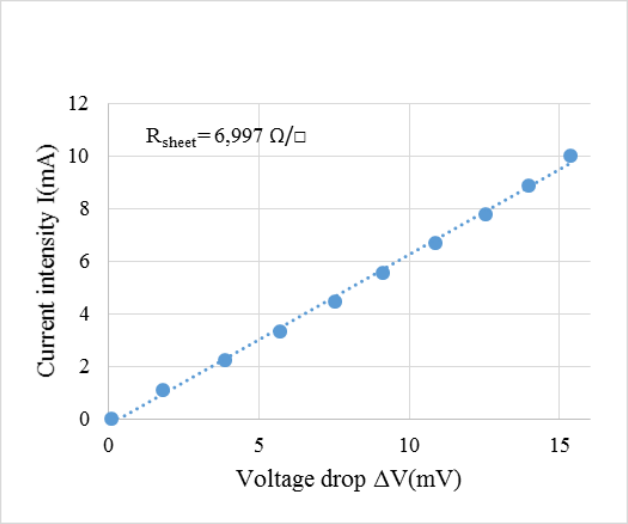


Figure 9: Variation of I(mA) as function of ΔV(mV) of silicon nanocones fabricated by etching 5-minutes processed SiNWs.

5 Conclusion

Silicon nanostructures (SiNWs and SiNCs) have been fabricated by metal-assisted chemical ething. The SEM characterizations have shown that SiNWs fabricated for 5 minutes are vertically straighter than the others. As a result, they exhibit lower reflectance (around 6%) in the 400-1100 nm wavelength range. The SiNCs were also fabricated by etching 5-minutes processed SiNWs. Comparated with SiNWs, SiNCs give the best antireflective properties. They allow to reduce the reflection down to 3% in the 350-1050 nm wavelength range. Measurements and comparison of sheet resistances from bulk silicon and silicon nanostructures, have demonstrated an improvement in the property of surface electrical conductivity.
We expect that with high quality surface passivation, performance of solar cells based on these fabricated silicon nanostructures can be greatly enhanced.

Acknowledgment This work was carried out with the support of the CEA-MITIC (Centre dExcellence Africain en Mathmatique, Informatique et TIC).

References

[1] Bilan du Forum Tchernobyl, organis par l'AIEA, l'OMS et le PNUD, septembre 2005

[2] S. Alexander et Reuters, Fukushima evacuation has killed more than earthquake and tsunami, survey says, NBC News.com, 10 septembre 2013.

[3] M.A. Green, Y. Hishikawa, W. Warta, et al, Solar cell efficiency tables (version 50), Progress in Photovoltaics 25, 2017, 668-676

[4] V. K. Sethi, M. Pandey and P. Shukla, Cost Boundary in Silicon Solar Panel, International Journal of Chemical Engineering and Applications, vol. 2, pp. 372-375, 2011.

[5] Bo Hua, Qingfeng Lin, Qianpeng Zhang and Zhiyong Fan, Efficient photon management with nanostructures for photovoltaics, Nanoscale, 5 2013, 66276640

[6] H. Wang, K. Lai, Y. Lin, C. Lin, and J. He, Periodic Si nanopillars arrays fabricated by colloidal lithography and catalytic etching for broadband and omnidirectional elimination of Fresnel reflection, Langmuir 26 (15), 2010, 12855-12858

[7] J. Lui, M. Ashmkhan, B. Wang, and F. Yi, Fabrication and reflection properties of silicon nanopillars by cesium chloride self-assembly and dry etching, Applied Surface Science 258, 2012, 8825-8830

[8] D. Dimova Malinovska, M. Sendova Vassileva, N. Tzenov, M. Kamenova, Preparation of thin porous silicon layers by stain etching, Thin solid films, vol. 297, pp. 9-12, 1997.

[9] X. Li, Metal assisted chemical etching for high aspect ratio nanostructures: A review of characteristics and applications in photovoltaics, Current Opinion in Solid State and Materials Science 16, 2012, 7181

[10] K. Peng, Y. Yan, S. Gao, and J. Zhu, Synthesis of Large-Area Silicon Nanowire Arrays via Self-Assembling Nanoelectrochemistry, Adv. Mater. No. 16, 2002, 1164-1167

[11] K Peng, Y Yan, S Gao, and J Zhu, Dendrite-assisted grow of silicon nanowires in electroless metal deposition, Advanced functional materials, Vol. 13, No. 2, 2003, 127-132

[12] T. Shimizu, N. Tanaka, Y/ Tada, Y. Hara, N. Nakamura, J. Taniuchi, K. Takase, T. Ito, and S. Shingubara, Fabrication of nanocone arrays by two step metal assisted chemical etching method, Microelectronic Engineering 153, 2016, 5559

[13] Z. Huang, N. Geyer, P. Werner, J. De Boor, and U. Gsele, Metal-Assisted Chemical Etching of Silicon: A Review, Adv. Mater., 23, pp. 285-308, 2011

[14] C. A. Bishop, Process Diagnostics and Coating Characteristics, Vacuum Deposition onto Webs, Films and Foils, 2015

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 45-64 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

W.A.S.T.E. R.E.D.U.C.E.: Waste Auditing Sensor Technology to Enhance the Reduction of Edible Discards in University Cafeterias & Eateries

Shiree Hughes*, Jiannan Zhai, Jason Hallstrom

Institute for Sensing and Embedded Network Systems Engineering, Florida Atlantic University, 33434, USA

ARTICLEINFO

Article history:
Received: 21 December, 2018
Accepted: 08 February, 2019
Online: 15 March, 2019

Keywords:
Internet of Things
Ambient displays
Food waste

ABSTRACT

Food waste disposal and food availability are both problems in the world today. The ability to monitor food waste is currently limited by the use of manual data collection and sparsity of monitoring. Using sensors to automate the process of measuring and collecting data on food waste would allow constant monitoring and provide a better dataset for analysis. This data can also be used to create ambient displays geared educate a target audience and produce behavioral changes. Student cafeterias serve as the focal point for data collection within this study, presenting both an abundance of data and unique challenges. We have created, deployed, and tested an autonomous, real-time system to collect data on food waste which includes an ambient display tailored to effect behavioral change in students.

1 Introduction

This paper is an extension of the work presented in *W.A.S.T.E. R.E.D.U.C.E.* [1].

In the U.S., approximately 40% of the food we cultivate goes to waste; yet, 1 in 6 people suffers from inconsistent access to food [2]. The cost of this waste is $165 billion per year, plus a large portion of methane emissions from U.S. landfills. Reducing the amount we waste by just 15% would generate enough food to feed approximately 25 million people, which would reduce food shortage problems by almost 50%[1]. It is imperative that we begin the task of educating society and encouraging changes in behavior.

Food waste studies are categorized based on where in the pipeline food is lost. Pre-consumer food waste is food that is discarded before reaching the end consumer; this includes food that is spoiled, dropped on the floor, or otherwise improperly handled. Post-consumer food waste consists of food that is prepared for and taken by a customer for consumption, but is eventually discarded; this would include leftover food on a customer's plate [2].

Educational institutions house future leaders; they are a desirable target to begin affecting change in food waste behavior. Campus cafeterias are typically the primary providers of meals to students, faculty, staff, and other visitors on college campuses across the nation. This makes them an ideal location for sampling post-consumer waste to get an accurate representation of food waste on a campus. As such, our first case study occurs in the main cafeteria on the Boca Raton campus of Florida Atlantic University (FAU), which services all students who have meal plans, plus drop-ins. Three different displays are introduced in the case study; each display conveys the amount of food wasted in real-time. A second case study was realized with one of the aforementioned displays in the cafeteria at A.D. Henderson University School, a laboratory school located on FAU's campus, which services elementary and middle school students. This location allows us to compare the effects of the display on different age ranges. We are also able to witness the effects of policies within different cafeterias that may aid (or hinder) food waste reduction.

Like many college cafeterias, FAU's dining hall is designed to be a la carte, allowing customers to pick and choose which sides and entrées they would like, including the choice of mixing and matching from the menu options. By allowing customers to choose their dishes, one might expect that it would reduce waste (e.g., a customer could request only the fries from a

*Shiree Hughes, 777 Glades Rd, Boca Raton, FL 33434, shughes2015@fau.edu

[1] This estimation was calculated using the current population of 324,121,843, as of July 31, 2016 [3].

https://dx.doi.org/10.25046/aj040207

burger-and-fries meal, which would eliminate wasting a burger). Yet, it is also possible that a customer may request a particular combination of dishes, be unsatisfied, and revisit the line for a more appetizing selection, leading the customer to waste the first plate. It is also possible that even if satisfied with the selection, the customer may still waste some of the food. U.S. portion sizes have been shown to be as much as 700% larger than the U.S. Department of Agriculture (USDA) recommendation, and as much as 263% larger than the Food and Drug Administration (FDA) recommendation [4]. While most research cites this as a contributing factor to obesity, it may also be a contributing factor to wastefulness. Information calling attention to this fact may caution students to ask for smaller portions.

Food waste audits are implemented in cafeterias of many institutions to generate information on how much food is wasted in their cafeterias for similar reasons to those noted above. However, most of these institutions only collect data for a short span of time, do not necessarily provide resolutions for reducing the waste that is found, and the process is not automated. An overview of food audits performed by institutions comparable to FAU can be found in Section 2. The purpose of WASTE REDUCE is to provide: (i) sustainable and on-going measurements of food waste; and (ii) visual cues designed to decrease the amount each customer wastes by encouraging more thoughtful choices.

The implementation of such a system has several challenges, related to both the implementation of the network and the visual cues used to evoke behavioral change. Cafeterias are dynamic environments and thus require a network that is able to withstand and/or adapt to sudden changes. Network reliability is a critical component to keeping the system operational and requires the network to adapt to communication issues such as interference or obstruction. To truly be an autonomous system, we must also ensure cafeteria employees are not required to change their behavior to use the system. This requires a robust design that withstands mishandling and minor leaks from garbage bags. The sensors must be power efficient, both for ease of use and for reliability. The devices within the network must also be compact enough to fit within a trash can. Most importantly, the sensors must be able to provide accurate, real-time data to create real-time feedback for the visual cues. The visual cues, which are implemented as a series of displays, must be seen by the target audience and the audience must be able to connect the information presented with their own behavior for change to take place. In addition, a display must resonate with the target audience to evoke change. After addressing each of these challenges, we designed and evaluated the use of three different displays with the system. Two of these displays were proven to evoke behavioral change in the target audience and reduced the amount of waste generated.

We present the design and implementation of **W**aste **A**uditing **S**ensor **T**echnology to **E**nhance the **R**eduction of **E**dible **D**iscards in **U**niversity **C**afeterias & **E**ateries (WASTE REDUCE), an automated method of measuring food waste and providing real-time feedback. We evaluate the system through a preliminary deployment and case studies. We also present the design and implementation of three ambient displays to provide visual cues to a targeted audience based on the real-time data provided through WASTE REDUCE. We evaluate the effectiveness of these displays on behavioral change through two case studies. [1].

Section 2 presents background and related work. Section 3 details our approach, including the hardware components, software implementation, and challenges for each component. Section 4 presents the results and evaluation of the case studies used to assess the effectiveness of the ambient display component of WASTE REDUCE and the associated displays. Section 5 concludes the work.

2 Background and Related Work

To better understand the design and implementation of each focus, we must first explore the background and related work concerning each area.

2.1 Campus Food Waste Auditing

Food waste is a major concern for society and many university cafeterias are engaging the issue through food waste audits. Food waste audits allow the cafeteria to collect sample data on how much food is being wasted in their dining halls and may or may not include campaigns to reduce this waste. Currently, these food waste audits are implemented manually. Highlighted below are methods and results from audits conducted by several universities around the country.

The University of California at Davis (UC-Davis) has an enrollment of approximately 35,000 students [5]. Using quarterly waste audits, UC-Davis saw a 30% decrease in food waste from 2012 to 2015. In the Fall of 2009, UC-Davis cafeterias saw an average of 3.6 oz (102.06 g) of waste per guest; this dropped to 1.6 oz (45.36 g) per guest by Spring of 2015. The primary goal of their food waste program is to have zero waste by 2020. A food waste audit is conducted in each of the university's cafeterias during lunch hours over 3 consecutive days each quarter. Part of the audit includes a visual display of the amount of food wasted, which is created by asking students to divide their waste into edible (e.g., leftover breadsticks), inedible (e.g., banana peels), and other (e.g., napkins). The edible food is then placed on display for guests to view and discuss. Other solutions implemented by UC-Davis include allowing students to ask for smaller portions, distributing samples for taste-testing, and donating unused food through the Food Recovery Network [6, 7]. (FAU also uses the Food Recovery Network to donate leftover food.) Since these audits only occur briefly during the semester and during one meal time, they can only provide a snapshot of the amount of waste

generated. This snapshot may not reflect the average amount of daily waste generated.

Georgia Institute of Technology (GT) has an enrollment of approximately 25,000 [8]. GT uses the LeanPath tracking system in their cafeteria to monitor and inform staff about food waste [9]. LeanPath is a commercial system that allows kitchen staff to record information on waste, such as food type, weight, and reason for disposal, each time something is thrown away. This system measures pre-consumer food waste, and provides detailed graphs and reports, estimates on cost, and email alerts for disposed amounts valued over a set threshold [10]. While the tool increased awareness among staff members, LeanPath requires extra effort from each staff member. To generate a report, LeanPath requires a staff member to weigh the food they are throwing away and provide additional information, such as his or her name, the type of food (e.g., dairy), what the food is (e.g., yogurt), reason for disposal (e.g., expired), and what type of container is being disposed (e.g., carton). This method has four main drawbacks: (i) it is an extra burden for staff who may have their hands occupied with food preparation; (ii) staff can easily forget or neglect to record items; (iii) it introduces a higher likelihood of human error; and (iv) it does not include post-consumer waste.

Michigan State University (MSU) has an enrollment of approximately 50,000 students [11]. MSU's research on food waste focused on cafeteria food waste in addition to off-campus food waste. Researchers discovered that off-campus, the average amount of food wasted was approximately 1.54 lbs (0.7 kg) per week for a single person. Using the number of weeks students are in school, compounded by the number of students living off-campus, researchers scaled their findings to estimate a total of 3,669,120 lbs (1,664,284.8 kg) of food waste per year from off-campus housing. In addition, MSU found that approximately 14,131 lbs (6,409.7 kg) of food was wasted per day in their cafeterias, with over 9,500 lbs (4,309.1 kg) of this weight associated with post-consumer waste [12]. Solutions to help reduce waste were printed along with these findings in a brochure for students. Recommendations were also given to the students on how to handle waste, including information on composting, recycling, and local produce. In addition, MSU created a food waste program that implements composting, and a student organic farm to help students handle food waste [13].

Florida International University (FIU) serves approximately 45,000 students [14]. FIU conducted a general waste audit and found that on average, 25.6% of their waste, and 4.5% of their recyclables were from food waste. For the audit, researchers collected samples of general waste across campus, sorting it by type and location. Food waste was one of the top three contributors to waste in each of the locations, and the primary contributor in housing and auxiliary areas. However, only one of their fourteen proposed solutions to waste reduction focused on food waste [14]. There was no clear campaign defined in the report to inform students of the audit results or to reduce waste generated.

The University of Miami (UM) has an enrollment of approximately 17,000 students [15]. Like FAU, UM uses tray-less dining halls to reduce the amount of food students take to the table. UM also reduced the amount of solid waste it generates by using a bio-digester to convert pre-consumer food waste to liquid. While UM does have a composting program on its campus, it is not integrated with their food waste management program [16]. UM reports the total processed waste generated by the campus as part of their Green U website, but they do not specifically report food waste [17].

2.2 Data Analysis and Feedback

Extensive research has been conducted in data presentation and visualization. Behavioral change cannot be realized without effectively conveying data to the target audience–in this case, cafeteria customers. Effectively conveying data requires guarding against habituation, presenting the data in compliance with known human-computer interaction (HCI) standards, and creating a lasting impression that encourages retainment of the information.

Habituation is the "diminishing of an innate response to a frequently repeated stimulus" [18]. Habituation is frequently used in post traumatic stress recovery to desensitize patients to images or thoughts that trigger their discomfort [19]. Repetitive exposure to images trains the brain to normalize the discomfort and anxiety provoked by the image. This is the opposite of what is desired for this project. Cafeteria customers should get a renewed sense of urgency each time they view the feedback–large quantities of food waste should not be normalized. Varied presentations of data will be used to avoid habituation.

HCI research has generated a wealth of knowledge pertaining to engaging customers with information displayed via a monitor. Adhering to these principles is crucial in designing a display that is successful in capturing cafeteria customers' attention. Lin et al. [20] outline eight criteria that should be taken into account for optimal engagement: stimulus-response compatibility, consistency, flexibility, learnability, minimal action, minimal memory load, perceptual limitation, and user guidance. Stimulus-response *compatibility* occurs when users have faster and more accurate responses for a set of attributes because the attributes naturally correspond. An example would be using red to highlight an error in the system versus using green; a user would naturally attribute red to a problem, whereas green would naturally lead the user to assume that everything was fine. *Consistency*, both within the system and across different systems improves usability. This goes hand-in-hand with compatibility; the use of consistency creates a natural correspondence among certain symbols, icons, layouts, etc. *Flexibility* enables a system to adapt to users' needs. This can range from allowing a user to customize the interface (e.g., changing the color scheme) to adapting to unique environments automatically (e.g., website designs that work for PCs,

tablets, and mobile phones). *Learnability* refers to the ease with which a person can learn to navigate an interface. The easier it is for a person to understand the format, the greater the user's satisfaction. *Minimal action* echoes the principle of learnability in the sense that users prefer interfaces that are straightforward and require minimal effort to complete a desired task. *Minimal memory load* addresses the working memory load required of the user to use the interface. Users perform better under minimal working memory loads. *Perceptual limitation* addresses the short-comings of human perception, such as color resolution. Designs should group like elements and use spatial and color resolution to ensure all details can be understood. *User guidance* refers to implementing cues and documentation, which gives the user information on how to use the system. Implemented correctly, user guidance aids with learnabiltiy and enabling a minimal memory load.

Creating a lasting impression in customers requires the information to be presented in a manner that achieves impact. Heath and Heath [21] outline six principles meant to enhance the "stickiness" of an idea. They define a "sticky" idea as one that remains with people over time. The six principles, which they refer to as SUCCES, are simplicity, unexpectedness, concreteness, credibility, emotions, and stories. *Simple* ideas are easier to remember than complex ideas; the authors argue that to make an idea stick in someone's mind, you have to reveal the core idea in simple terms. *Unexpected* information and surprises catch people off guard, which sparks curiosity and causes them to think about what they are being told. *Concreteness* refers to presenting clear ideas. Information that is abstract to the point of ambiguity is not likely to achieve impact. *Credibility* is needed to convince an audience that the information presented is accurate and reliable. *Emotions* have a way of controlling people when logic fails–in many cases invoking strong feelings of sympathy, regret, or anger. Heath and Heath give the example of persuading people to donate to specific people who are in need, which has a higher success rate than persuading people to donate to an anonymous group of people in need. *Stories* provide a plot and direction, which is often easier to remember than bulleted lists. Heath and Heath concede that each idea might not lend itself to every principle, but suggest that employing a number of these ideas will improve impact. We employ these techniques in the design of each display for WASTE REDUCE to ensure the magnitude of food wasted "sticks" in consumers minds to prompt a change in their behavior.

Stasko et al. [22] explore the technique of conveying information through art. The authors focus on ambient, peripheral displays. The displays are designed to smoothly adapt based on the underlying information being presented, while maintaining an aesthetically appealing appearance. Non-critical information, such as information on food waste, is usually the driving source of such displays. Critical information, such as a medical alert, is not ideal for these displays because the goal is to produce an informative, yet "calm," tool. The aim is to reduce cognitive disruptions so that seeking out the information does not require the user to completely shift his or her train of thought. Maintaining the calmness of the display allows the display to move from a user's periphery to focal attention, and then back to the periphery. Ambient displays began with the emergence of Dangling String [23], which conveyed information on how busy a network was by causing a dangling string to twitch each time bits of data were transmitted across the network.

Ariely [24] explores human behavior, specifically that of customers, in relation to information control. Information control refers to the ability of a user to control what information they receive. Research shows that there are both advantages and disadvantages to restricting users' control over the information presented. A key advantage is that users are able to choose topics that interest them the most. The key disadvantage is that this increases cognitive burden since users must configure what information is presented. In Ariely's studies, users who were able to choose the way in which information was presented learned the information better. Results suggest an improvement in users' decision making skills when information control is used.

2.3 Wireless Sensor Networks

Wireless sensor networks are formed by several sensors connected via wireless communication protocols [25, 26]. These networks allow specific information to be collected and transmitted to a central location for monitoring or analysis. There are a wide variety of sensors available for use, each able to provide specific information about an environment. A representative example of a sensor is a load cell, which can be used to determine the weight of an object. Such sensors can be used independently or in conjunction with other sensors to determine pertinent information about the environment in which the sensors are deployed (such as the weight of objects placed in a container, or water quality in a river). When connected to a wireless communication device, the information can be transmitted within the network. Networks usually have a *base station*, which receives the data transmitted by the sensors in the network, and then stores the information in a central location, such as a server. Users can then access the data without needing to physically interact with the environment. This is ideal for monitoring information in unpleasant locations, such as a garbage can, and for long-term monitoring, where it is infeasible for a person to be continually present to collect the data.

The sensors and radio are connected and controlled by a microcontroller. The microcontroller serves as a centralized controller and uses software modules, commonly known as drivers, to control the sensors, radio, and other peripherals. The overall device is typically small, embedded in the environment being monitored, and battery powered. As such, power and memory are

typically limited [27]. To maximize the longevity of the battery, software must be written to take advantage of low-power and sleep modes for sensors and microcontrollers. To satisfy the memory constraints software must also be concise. A few example devices that could exist in a sensor network include the RaspberryPi [28], Aurdino [29], and BeagleBone Black [30].

Connectivity is a primary concern with sensor networks. Data is typically transmitted through small wireless packets. Packets contain additional information for the destination device, such as the source of the data, and a checksum to verify that the message has not been corrupted. Communication protocols must be robust enough to handle interference and failures within a network. The quality of links within the network, which is a predictor of system performance, is often analyzed to debug or enhance communication within a network. Two common methods of defining network quality are Received Signal Strength Indication (RSSI) [31] and packet reception rate. RSSI provides a measurement of the strength of a received transmission. Packet reception rate conveys the percentage of packets successfully transmitted with respect to the total number of packets transmitted. The higher the percentage, the better the link quality. We use RSSI and packet reception rate to examine the quality of the links within WASTE REDUCE.

2.4 Data-Driven Visualization

WASTE REDUCE is concerned with affecting behavioral change through dynamic displays, driven by automated data collection.

Current methods for collecting food waste data and presenting information for social norming campaigns, as discussed previously, involve manual collection of data and the use of static advertising material. Manual collection of data not only introduces human error—what if someone forgets to weigh the waste before emptying it?—but is difficult to sustain, as witnessed in the audits discussed previously. Static advertising doesn't account for changes in the data and risks habituation. Data collected over a short interval of time may or may not be indicative of normal behavior, and since data is only collected during the campaign, it is difficult to assess if the campaign was successful in raising awareness or reducing waste. This also makes it difficult to provide students with feedback about their progress in reducing waste. Campaigns are left with the option of either disseminating the same information until the next audit, which is likely to result in habituation, or place the responsibility of continuing the message on the students, who would need to remember and gauge their own behavior.

The cognitive burden of a target audience can be managed by presenting data to an audience in an appealing and informative manner. Since data-driven visualizations are controlled by data inputs, the display is more readily updated to reflect data changes or needs of the target audience. Data-driven displays can be used to analyze real-time data and change on the fly.

Plaue et al. [32] explore the efficacy of presenting information through peripheral displays through *InfoCanvas*. InfoCanvas is a form of information art in which the art is manipulated to communicate information. The authors chose 10 specific pieces of information, generally found on a home page, to present in a scene depicted by InfoCanvas. Each piece of information was mapped to an image within the picture, and the image changed based upon the information. For example, airfare prices were represented by a kite flying in the sky; when the kite is closer to the horizon, it signifies a low price, and when the kite is high in the sky, it signifies a higher price. Plaue et al. [32] compare users' ability to remember the information using the abstract InfoCanvas versus a text-only display and a typical web interface (containing a mix of text and graphics). After allowing users to view a display for 8 seconds, users were given a recall test to assess how well the information was retained. For 7 of the 10 elements of information, InfoCanvas had the highest rate of recall. Users were also given a survey to rank the 3 displays based on ease of information recall, effectiveness of data presented, and visual appeal. Users rated InfoCanvas highest for both ease of information recall and visual appeal. Some users expressed a preference for the web portal display citing it to be "more professional," "logical and precise," while others expressed a preference for InfoCanvas. Those who preferred the web portal and those who preferred InfoCanvas mentioned the need for adaptation to the InfoCanvas display.

Dahley et al. [33] compare the use of ambient displays to the cues we pick up in nature. Drawing inspiration from nature, the authors created *Water Lamp* and *Pinwheels*. Water Lamp is inspired by the creation of ripples from raindrops. Using data to simulate raindrops, computer-controlled solenoids are used to create ripples in a tray filled with water. A light shines from beneath the tray, displaying the ripple effect through light and shadow on the ceiling above. Water Lamp can be driven from multiple sources of information, enabling the display to behave as an indicator of the weather outside or a white-noise effect. Pinwheels was created to convey information on air flow. Air flow data is represented using pulse width modulation to control a motor, which spins a pinwheel at a speed reflective of air flow. Both devices function as a display and a physical reminder of the associated data. Identifying a physical representation that would convey information on food waste is a difficult task and not likely to be intuitive to our audience; we limit our display to the virtual world.

Mankoff et al. [34] create BusMobile and Daylight Display to evaluate the effectiveness of ambient displays. Both displays were created to provide information to college students in a window-less lab. The displays were created after surveying the students on which information was of most interest to them. BusMobile uses information from the published local bus schedules and the current time to inform students how

near or far a bus is to its scheduled stop. Each bus is represented by a "bus token," which moves to the top of the display as it approaches the bus stop. Daylight Display was designed to inform students about outdoor lighting. A lamp was modified to brighten and dim according to the expected amount of light outside for the current time of day. Both displays use static data and will not be accurate if there is a deviation from the norm– e.g., a bus stuck in traffic, or a cloudy day when it becomes dark earlier than expected. Surveys from students revealed the Daylight Display was less effective than the BusMobile; many students assumed the lamp was broken and did not notice information being conveyed.

Heiner [35] et al. describe Information Percolator, a display created from tubes filled with water. Bubbles in the water are produced to behave as pixels, allowing the creation of an image across the tubes. One application for the device is a clock application. The clock application is designed to sync with a user's calendar and change the display to notify the user of an event. Information Percolator can also display specific information, such as the availability of a co-worker or current stock market trends. Heiner [35] et al. present a new and interesting way to convey information; however, the effectiveness of this method was not explored.

Snapshot [36] was developed to introduce data-driven visualization techniques to the game of hockey. Snapshot marries current hockey analytics with visualization and is designed to give both teams and fans a clearer picture of patterns within the game. In hockey, teams may be concerned with where on the court shots are most frequently taken by their opponent. To visualize this data, Snapshot uses a monochromatic heat map of the rink, where variations in color are determined by opacity. The opacity of a particular location is set based on the frequency of shots in the region with respect to the number of shots taken overall and the size of the region. Snapshot also provides visualization of shot distance using a radial heat map. Shots are divided into bins based on distance and mapped to rings of distance around the goal; each ring is colored based on the number of shots that fall in that ring. To determine the effectiveness of such a technique, Snapshot was evaluated by 3 professional hockey analysts. Each analyst discovered something unique while using the tool. The first analyst was able to pick up on patterns in multiple rinks that were previously thought to exist only in a particular rink. The second analyst found patterns related to how teams shoot and score at home versus away. The final analyst used Snapshot to investigate "sweet spots" in the rink and found that his hypothesis of shooting from a goalie's weak side did not yield more goals. All 3 analysts expressed interest in using the tool beyond the evaluation period to dig deeper into their theories. While Snapshot is not a real-time tool and is specific to hockey, it accomplishes our basic goal of conveying accurate data in a manner that stimulates interest.

Current solutions do not address the challenges associated with creating a network to monitor food waste data, nor do current ambient display techniques attempt to work in conjunction with a wireless sensor network to change behavior. The focus of ambient display research is on informing an audience rather than affecting the behavior of the audience. We present novel approaches to specific problems which were previously unaddressed.

3 Approach

We accomplish automated collection and analysis of food waste data through an accurate, robust, and efficient wireless sensor network. The method used to determine the amount of food wasted must be able to accurately detect small changes in weight for the system to provide a complete picture of waste accumulation over time. The system must be robust enough to maintain functionality despite unexpected events, such as mishandling of trash cans or network interference. Efficiency is necessary to minimize power consumption which will extend the lifetime of the battery and maintain functionality over long time periods. The device must also be compact enough to fit inside the trash cans used by the cafeteria.

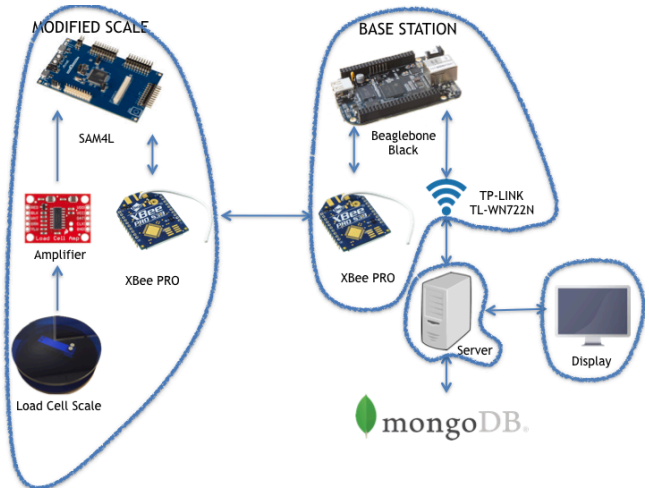


Figure 1: Food Waste Monitoring Architecture

3.1 Hardware

The system architecture, shown in Figure 1, consists of four main components: a modified scale (located within each trash can), a base station, a back-end server, and a front-end display.

3.1.1 Modified Scale

The modified scale was created using a 10kg wide-bar load cell [37], an HX711 [38] amplifier, an XBee-PRO 900HP S3B [39] radio, and a SAM4L Xplained Pro [40] board. The SAM4L Xplained Pro is a prototyping board for Atmel's SAM4L Cortex™-M4 based microcontroller. The load cell translates force exerted on the scale to an electrical signal, which is amplified by the HX711. This raw value is transmitted by the

XBee radio to a base station. A PCB board was designed for the XBee-PRO and HX711 to plug into the SAM4L Xplained Pro extension header. This enabled a compact, 4" x 5" design conducive to the limited space within each trash can.

3.1.2 Base Station

The base station is driven by a BeagleBone Black [30], which is connected to an XBee-PRO 900HP S3B [39] radio and a TL-WN722N 150 Mbps high-gain WiFi dongle [41]. A Java program is used to receive packets transmitted from the radios on each device, determine which device sent each packet, convert the payload of the packet to pounds, and transmit the human-readable information to a server.

3.1.3 Server

Collected data is stored in MongoDB [42] so that it may be queried for later use. Since the server is protected by a firewall and cannot be accessed remotely, we adopted a RESTful design. The RESTful design allows access to the data through text strings which can be passed from the server to the front-end. We developed a server-side function, called using HTTP-GET requests. The function queries MongoDB based on parameters supplied by the front-end and returns a JSON string containing the data.

3.1.4 Design Challenges

Detailed discussion of design challenges discovered and overcome during the initial deployment can be found in the original publication [43]. These challenges included both situations in our control, such as the antenna strength of the radios or reliability of the communication protocol, and out of our control, such as the reliability of the WiFi network in the deployment location or how the devices are handled.

3.2 Software

3.2.1 Modified Scale, SAM4L Implementation

Figure 2 illustrates the flow for the implementation of the modified scale. First, the SAM4L and all connected sensors are initialized. After initialization, the same series of steps are repeated for each cycle. The microcontroller and radio are brought to their active states. A sample is then taken. Once the sample is taken, local time is attached to the sample. Next, a packet containing the sample and necessary timestamp information is transmitted to the base station. Local time on the scale at the time of sampling and at the time of transmission is kept; therefore, if the network is experiencing interference, and it takes multiple retries to successfully transmit the data, delays that occurred between the time of sampling and the successful transmission to the base station can be accounted for. The SAM4L marks local time via an asynchronous timer, driven by a 32.767kHz crystal. The timer updates a counter which defines the local time in number of ticks since boot. This information is used to convert from local time to wall clock time at the base station, achieving time synchronization. If no acknowledgment is received from the base station, it is assumed the transmission was unsuccessful, and the sample is placed in a queue to be retransmitted during the next cycle. If an acknowledgment is received, the transmission is considered successful and the devices are put into sleep mode for 60 seconds before restarting the cycle. Placing the radio and microcontroller to sleep between cycles reduces power consumption. This is crucial since it is undesirable to change the batteries frequently. An analysis of power consumption can be found in Section 3.3.

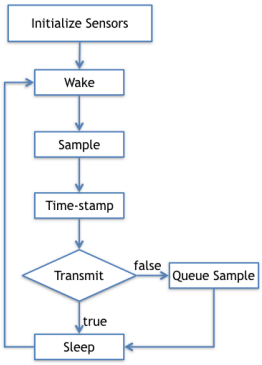


Figure 2: Modified Scale Implementation (SAM4L)

3.2.2 BeagleBone Implementation

Figure 3 depicts the flow of execution for the base station. First, the BeagleBone Black and XBee radio are initialized. Once initialized, the radio listens for packets from the modified scale. Once a packet is received, the local time from the modified scale is converted to wall clock time. Next, the base station transmits an acknowledgment to the scale to indicate receipt of the data. The scale then converts the sample from raw data to pounds. The final sample is then transmitted to the server over WiFi. If the sample cannot be successfully transmitted to the server, it is placed in a queue and the transmission is retried at a later time. After attempting the transmission, the radio once again listens for packets.

Time Conversion. Figure 4 depicts the process of timestamp conversion. The SAM4L marks local time using a counter driven by a 32.767kHz oscillator crystal. After a prescaler is applied, the effective increment of the clock is 1 millisecond. The counter is updated each time there is an overflow. From this, we know each counter increment represents 32,767 milliseconds. `transmissionTime` is approximately equivalent to the wall clock time at reception. Given this information, we can identify the local time at which the sample was taken (the difference between `transmissionTime` and `sampleTime`) and convert this into milliseconds elapsed. We can then add those milliseconds to the

current wall clock time to estimate the wall clock equivalent of the sample time.

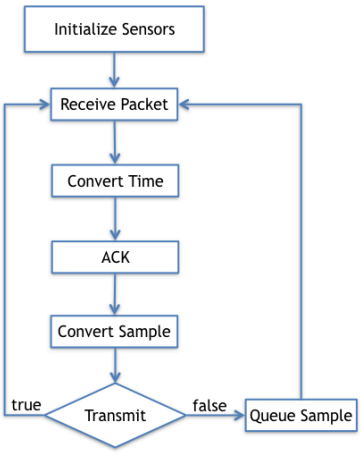


Figure 3: BeagleBone Black Implementation

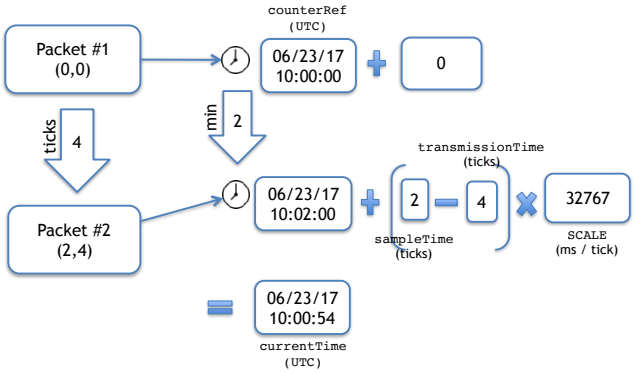


Figure 4: Time Conversion Flowchart

Raw Data Conversion. Equation 1 is used to convert raw data to pounds. The `offset` parameter is calculated by taking the average over 10 readings while nothing is placed on the scale. `scale` is calculated just prior to deployment using the calculated offset and an object of known weight. Averages over 10 readings are used[2], and the process is repeated with multiple objects of known weight to ensure the accuracy of the `scale` parameter.

$$knownWeight = (raw - offset)/scale \quad (1)$$

Tables 1 and 2 detail accuracy for each deployment. Accuracy of the scale varied negligibly between deployments due to differences in hardware mounting. Table 1 shows the accuracy, in pounds, for the case study carried out at Atlantic Dining Hall (Case Study #1). Table 2 shows the accuracy, in pounds, for the case study carried out at A.D. Henderson University School (Case Study #2). The objects of known weight were a power drill and a cell phone. Different cell phones were used between the pilots at the two schools. For the objects of known weight, and the case when the scale is empty, the tables show the average weight measured by the scale in each trash can followed by percent error. Percent error is calculated using Equation 2. As shown by the percent error, the scale is more accurate for heavier objects. This means we can be more confident as the amount of waste increases. We take this into consideration when identifying when the trash can was emptied; this is discussed in Section 3.4.1.

$$percent_error = (measured - actual)/actual \quad (2)$$

3.3 Power Consumption

Power consumption is an important factor in most wireless sensor networks and especially in this network. Replacing the scale's battery requires removing the top plate of the scale to gain access to the enclosure, all of which are located in the bottom of the trash can. We seek to maximize the lifetime of the battery by taking advantage of the sleep features of the XBee and SAM4L to minimize power consumption and limit inconvenience.

Figure 5 illustrates power consumption over time (x-axis) in mA (y-axis) for 15 cycles. We observe that power consumption is generally consistent across cycles. There are three observable states within each cycle: sleep, sample, and transmit. The lowest power consumption occurs at approximately 5mA, while both the SAM4L and XBee are in their respective sleep modes. Next, the SAM4L wakes and takes a sample. This appears as a defined step and shows a consumption of approximately 21mA. At the end of the cycle, the radio is awakened, and the sample is transmitted. Power consumption is not consistent during this state in the cycle, but averages 107mA.

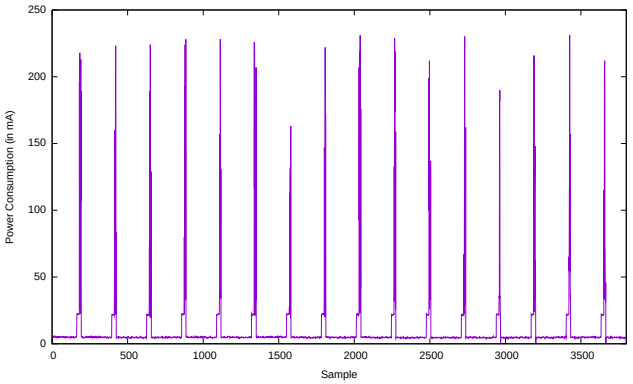


Figure 5: Power Consumption (in mA)

The average power consumption for each state is shown for 10 cycles in Figure 6. The x-axis shows the cycle, and the y-axis shows power consumption in mA. This confirms the variability of power consumed during transmission. We can estimate the average power consumption during each cycle by scaling the average power consumption of each state according to its duration and summing the values. In our case, the cycle is broken down as follows: 86% sleep, 8% sampling, and 6% transmission. Figure 7 shows the average power consumption for the 10 cycles. Again, the x-axis shows

[2]Each sample is based on 100 readings from the load cell as a result of the oversampling method discussed previously.

Table 1: Atlantic Dining Hall – Scale Accuracy (in lbs)

	Empty		Power Drill		Cell Phone	
	Measured	Error	Measured	Error	Measured	Error
Actual Weight	0.000	-	3.200	-	0.284	-
Trash Can #1	0.048	NaN	3.220	0.625%	0.345	21.479%
Trash Can #2	0.000	0.000%	3.200	0.000%	0.293	3.000%

Table 2: A.D. Henderson University School – Scale Accuracy (in lbs)

	Empty		Power Drill		Cell Phone	
	Measured	Error	Measured	Error	Measured	Error
Actual Weight	0.000	-	3.200	-	0.304	-
Trash Can #1	0.042	NaN	3.200	0.000%	0.359	18.092%
Trash Can #2	0.016	0.000%	3.199	-0.031%	0.336	10.526%

the cycle, and the y-axis shows power consumption in mA. On average, the current draw is 12mA. Using a 6000mAh battery the scale can be powered continuously for just under a month.

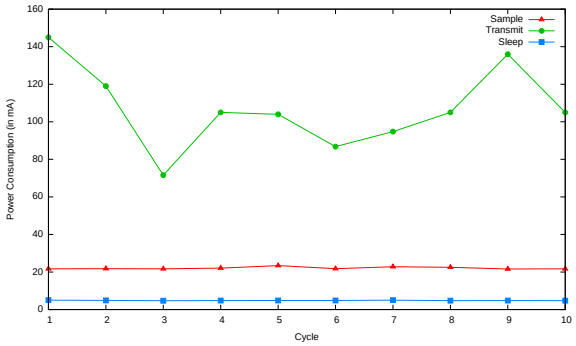


Figure 6: Average Power Consumption Per State

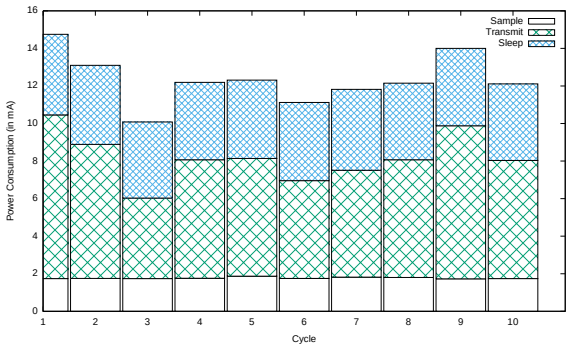


Figure 7: Power Consumption per Cycle by Active State

3.4 Front-End Display

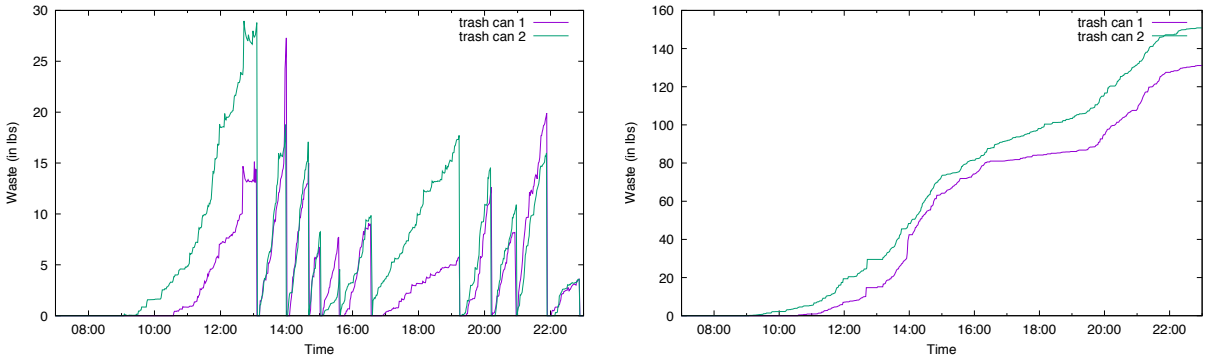
Three front-end displays were designed to test whether ambient displays could be used to change behavior. The front-end displays are JavaScript-based web applications. Graphs were created using the open source Data-Driven Documents (D3) library [44]. The application fetches information from the server using an HTTP-GET request and uses the data to create displays.

3.4.1 Calculation of Total Waste

Figure 8 illustrates two ways to present the waste measured over time. The first is used to analyze the time periods when the most waste is generated and the second is shown in the display to inform of overall waste. The x-axis shows the time of day the sample was taken, and the y-axis shows the waste, in pounds. Since the modified scale reports the weight of waste *currently* in the trash can (Figure 8(a)), a method has to be devised to approximate the *total* waste over time (Figure 8(b)). Total waste should be the sum of the local maxima, which should occur just before the waste is emptied. However, there are two scenarios that must be taken into account: (i) the scale may never register as empty, and (ii) there may be a spike in the data from someone pressing down on the trash bag or bumping the trash can.

Based on the accuracy of the scale, discussed in Section 3.2.2, an *empty* scale should be defined as a measurement below 0.05 lbs. However, during peak hours, it was observed that students may crowd around the waste area and empty waste in the trash can immediately after it has been emptied. When this occurs, the scale never registers a zero value. To account for this, when calculating the total waste, we check if the current sample is significantly less than the previous sample to identify if the waste has been emptied. *Significantly less* is defined as a difference of 10 lbs or more based on observations from the preliminary data.

Figure 8(a) illustrates the second scenario involving an anomalous reading, just before 12:00 in both trash cans. Since this occurs in both cans, the anomaly is likely from someone lifting the trash bag to tie it just before emptying the waste. (If pressure were applied to cause the spike, the corresponding area would have a concave shape.) In this case, the maximum reading before the empty event is added to the total. (In the opposite case, we take the value after the local maxima.) When the algorithm checks if the current sample is less than the previous sample, it checks for this scenario as well. Between each empty event, we save the maximum observed sampled. This value is compared to the current sample. If the current sample is more than 1 pound less than the maximum observed sample, but



(a) Measured Weight

(b) Accumulated Waste

Figure 8: Finding the Total Waste

not more than 10 lbs, the current sample is considered the real maximum. The previously observed maximum is then subtracted from the total waste and the current sample is added. The maximum observed sample is then set to the current sample. This method accommodates cases like the one shown in Figure 8(a), since the waste settles at its original weight before being emptied. The calculated accumulated waste for each trash can using this method is shown in Figure 8(b).

3.4.2 Display Setup

Figure 9 show the display setup for each location. Figure 9(a) [43] illustrates the setup inside Atlantic Dining Hall. This cafeteria has a single point of entry for students. Our display is mounted right as students enter, beside the pay stations. The display was placed here for ease of installation and maximum visibility. The optimal location for the display would likely be near the trash cans. This would allow students to make a direct connection to the information presented and the waste filling up the cans. The trash cans are not located near a power supply; therefore, it wasn't feasible to place a monitor there. A battery powered tablet would have been the only option for that location. However, the tablet would be too small to capture attention and would require recharging each day. Excluding the space directly beside the trash cans, the entry point near the pay station has the maximum visibility for customers. Everyone who enters the cafeteria must pass this location, and many are forced to linger in the vicinity as they wait in line to pay, providing more time to view the display. The displays are updated every minute so students can view the increase in waste in real-time.

The display for Case Study #2, at A.D. Henderson University School, was closer to the trash cans, as shown in Figure 9(b). However, we were still unable to place the display directly behind the cans as desired. Instead, the display was located in the front of the room, directly opposite to the trash cans.

A total of three displays are evaluated in Case Study #1. There are several reasons to test multiple displays. Each display focuses on one concept. Display 1 focuses on simplicity, Display 2 focuses on context, and Display 3 focuses on humanizing the issue. The variety of displays allows us to investigate if one method has the more impact than another. Ideally, if WASTE REDUCE were a permanent fixture in a cafeteria, multiple displays would be used in randomized rotation to avoid habituation.

All three displays use the same basic layout. The top of the screen displays the date and location. The left side of the display contains an image, which is altered by the data provided by WASTE REDUCE. The right side of the display consists of 3 graphs: (i) a gauge comparing the current day's total food waste accumulation in comparison to the previous week's maximum total food waste accumulation, (ii) a gauge comparing the current day's total food waste accumulation in comparison to the previous week's minimum total food waste accumulation, and (iii) a line graph showing accumulation over time for that day. This allows students to compare their progress in reducing waste. The gauges were updated for each display based on feedback from cafeteria employees about student confusion on what information the gauges represented and conveyed. The goal of conveying the max and min waste values for the week remains the same throughout each display. These differences are discussed in detail below. Below the two gauges is a line graph of time versus total weight in pounds. The total amount of waste, starting at breakfast, seen by each trash can is plotted on the graph; this graph is present for all displays.

For each display, we adapted 4 principles from Heath and Heath's [21] SUCCES method. Table 3 identifies which principles were applied to each display.

Table 3: Application of SUCCES Principles

	Display 1	Display 2	Display 3
Simple	✓	✓	
Unexpected	✓	✓	
Concrete	✓		✓
Credible	✓	✓	✓
Emotional		✓	✓
Stories			✓

(a) Atlantic Dining Hall (b) A.D. Henderson University School

Figure 9: Displays

3.4.3 Display 1

An example of the first display is shown in Figure 10 [43]. This display is the most straight-forward. It was designed to illustrate the message of food waste reduction in the simplest terms. Since reducing the cognitive load is one of the goals of an ambient display, and there was no opportunity to introduce or explain the display to the target audience, a no-frills approach was deemed best for the initial display.

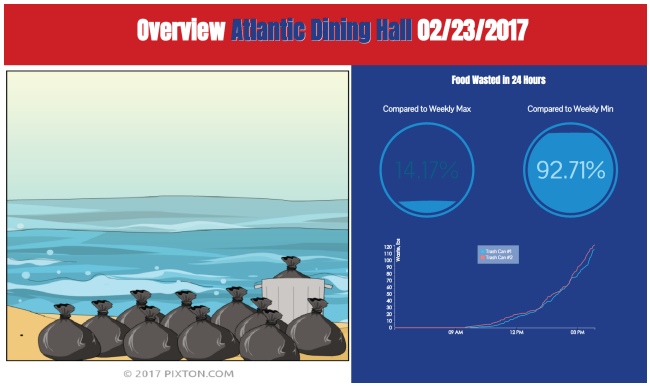


Figure 10: Display 1 for Case Study #1

The main section shows an image of the beach,[3] which becomes littered with trash bags in accordance with how much waste is generated. Preliminary data allowed us to estimate the average weight of the trash when emptied to be 20 lbs (9.07 kg). Therefore, for every 20 lbs wasted, a new trash bag is added to the beach scene in the display.

To the right of the main image is a sidebar of additional information. There are two gauges that show the amount of waste generated at present compared to the maximum and minimum observed waste over the previous week. The value shown is calculated by taking the current waste divided by the maximum and minimum, respectively. This allows us to present the current waste as a percentage of the highest and lowest waste observed the previous week. The gauges fill up according to percentage and the filled portion is animated to look like liquid.

3.4.4 Display 2

Figure 11 depicts a sample of the second display used for Case Study #1 and the only display used for Case Study #2. The display was designed to encourage customers to relate their waste to consumable food. Although this requires a slightly heavier cognitive burden than Display 1, since customers must connect the image of consumable food to wasted food, giving wasted food context in terms of consumable food is more likely to "stick" in consumer's minds. This display capitalizes on 4 out of 6 (67%) of Heath and Heath's [21] SUCCES method.

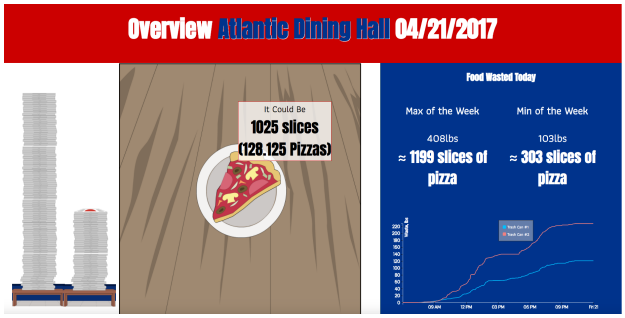


Figure 11: Display 2 for Case Study #1

Instead of using trash bags to depict how much was wasted, this display uses slices of pizza. Pizza was chosen as the food item to display because it is a common staple among college students.

The United States Department of Agriculture maintains a food composition database that provides the average weights of food items by serving size and brand [45]. Using this information, we determined the average weight of a slice of Papa John's *"Supreme Pizza"* to be approximately 0.34 lbs (.15 kg). We use Papa John's as the brand because the university has a contract with Papa John's; it is the pizza typically served on campus. In the pizza slice graphic used for

[3]The image was chosen due to the university's proximity to the ocean, and students' likelihood of relating to the image.

the display, the toppings present are similar to what is served on Papa John's Supreme Pizza.

A minor change to update the header of the display to reflect the new location was made before redeploying for Case Study #2. Since A.D. Henderson University School was too long to comfortably fit on one line, we simply called the page "Food Waste Overview."

Another change in the second display involved the gauges used for comparing weekly progress toward food waste reduction. The cafeteria staff expressed confusion, both among themselves and with the students, concerning what the original gauges were trying to convey. For clarification, the animated gauges were replaced with static text defining the maximum and minimum amount of waste generated for the week. These values were shown in both pounds and number of pizza slices.

3.4.5 Display 3

The main goal of Display 3 is to humanize the issue and remind customers that this issue is continuing in the background of our lives. Another goal for Display 3 is to avoid habituation. Although the other two displays update according to the data in real-time, the images will follow the same general pattern each day. This might cause students to lose interest in the display. Display 3 is guaranteed to have a different image each day, with the goal of preventing habituation and keeping customers interested in checking the display for new information.

Figure 12 depicts a sample of the third display for Case Study #1. There are two differences between this display and the previous displays. The focal image, featured to the left of the screen, depicts a story about 4 characters. The characters are designed to project diversity to ensure all customers feel included in the story. The story is driven by the amount of food wasted each day. Food waste for the day is compared to the average amount of food wasted over the previous week. When more food is wasted, the characters experience a bad day. When less food is wasted, the characters experience a positive day. As with the first display, trash bags appear in the scene to emphasize how much is currently being wasted.

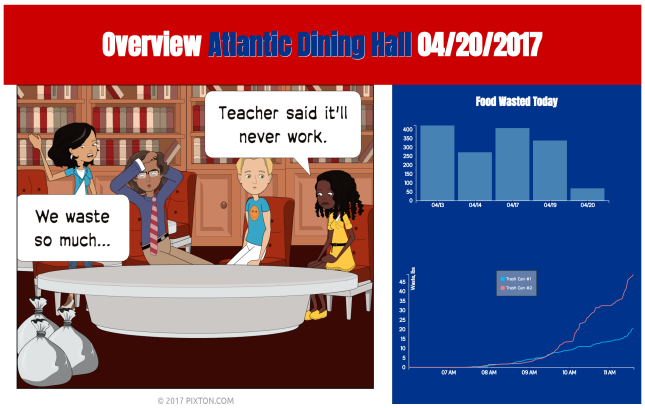


Figure 12: Display 3 for Case Study #1

Another change is in the depiction used to compare the maximum and minimum food wasted over the week. Since pizza slices were no longer being displayed, it no longer made sense to display weight in pizza slices as the maximum and minimum. Since the story follows the accumulation of waste each day, we thought a bar graph allowing customers to compare waste each day would be the most beneficial.

4 Case Studies

Two case studies were carried out to evaluate the efficacy of WASTE REDUCE to affect behavior change and reduce the amount of food wasted in a cafeteria setting. The first case study took place at a university cafeteria; it spanned 7 weeks and tested 3 displays. The second case study was carried out at a grade-school cafeteria over 2 weeks and tested 1 display. For each case study, we obtained a baseline of waste behavior during the first week.

4.1 Technical Difficulties

Although the system was thoroughly tested before beginning the case studies, there were still a few technical difficulties during the case studies. Otherwise, the data was collected over consecutive days and weeks.

4.1.1 Display 1

After the week of baseline data was collected, the cafeteria was closed a week for Spring Break. When the cafeteria reopened, there were technical difficulties the first two days and data could not be collected. This postponed the introduction of Display 1, but did not create a problem for the case study. Another technical difficulty occurred on the Monday during the second week the display was deployed (Week 3). The monitor that allows students to view the display experienced technical difficulties and was not visible until mid-afternoon for that day.

4.1.2 Display 3

Display 3 had one day of technical difficulties. The hotspot providing WiFi connectivity for the system malfunctioned several times during the Tuesday of the week this display was deployed (Week 6). The device could not maintain its connection to the cellular network; thus, the base station could not establish a WiFi connection. Data was not lost due to reboots of the device to troubleshoot and facilitate reconnection to the network.

4.2 Case Study #1

A case study was carried out in FAU's main campus dining hall to determine the efficacy of WASTE REDUCE. The data presented in this section was collected over seven weeks. The first week (Week 1) serves as a

baseline for comparing the effects of the ambient display, and thus, only the monitoring system was active during this week. During the second and third weeks (Week 2 and Week 3), the first ambient display was also present in the cafeteria. During the fourth and sixth weeks data was collected but no display was present. This was done to reduce lingering effects of the previous display when the next display was introduced. Display 2 was shown during the fifth week (Week 5), and Display 3 was shown during the final week (Week 7). Due to the cafeteria closing for Spring Break, there is actually one week gap between Week 1 and Week 2.

The campus cafeteria provided the number of meals sold each day during this time period. This information is used to normalize the data by determining waste per meal sold. The number of meals sold is the best estimate of how many people contributed to the waste. This allows better comparison of a day in which the cafeteria may have wasted 500 lbs while serving 5,000 people, versus a day when the cafeteria wasted 500 lbs, but only served 3,000 people. Since the number of meals sold per day is available only as a total, we can only normalize the total waste for the day, and not the real-time data throughout the day.

The cafeteria was closed the weekend before and after Spring Break; therefore, we could not establish a baseline for Saturday and Sunday. In addition, the cafeteria has restricted hours, a different meal service, and receives significantly less traffic during the weekends. Saturday and Sunday are not included in the analysis.

As traffic increases in the cafeteria, the number of trash cans used to contain waste varies, a factor that was unclear in the initial deployment. Only two instrumented cans were used during the case study; however, these cans were always present for student waste. More trash cans may have been present during various times within the study. Thus, all results represent a minimum bound on food wasted in the cafeteria.

4.2.1 Display 1

We will briefly summarize the results of Display 1 for the sake of comparison. Greater detail for this display can be found in the original publication [1].

To determine changes in waste behavior, we examined both total and normalized waste per day for the baseline week (Week 1) and the days the display was present (Week 2 and Week 3). Figure 13 illustrates these results. The x-axis shows the day of the week and the y-axis shows amount of waste (in pounds). For Figure 13(a), the waste is total waste for the day and for Figure 13(b), the waste is waste per meal.

The observation on Friday illustrates why comparing the total waste generated is not sufficient to determine behavioral change. In Figure 13(a), Friday is the only day in which total waste did not decrease after the introduction of the display. This divergence from the trend is because Friday of Week 1 was also the beginning of Spring Break; fewer people ate in the cafeteria that day, leading to less overall food waste. We normalize the day by diving the total waste observed by account the number of meals sold that day, shown in Figure 13(b). The normalized data confirms waste per meal decreased after the introduction of the display even on Friday.

We expected waste for Week 3, the second week the display was present, to continue to decrease or remain the same as Week 2 (the first week the display was present). Figure 13(b) shows a slight decrease for majority of the days in Week 3 compared to Week 2. The increase in waste on Monday of Week 3 compared to Monday of Week 2 is likely due to the technical difficulties experienced with the monitor that day as discussed previously.

Overall, These figures show a consistent decrease in waste per meal for each day the display was present.

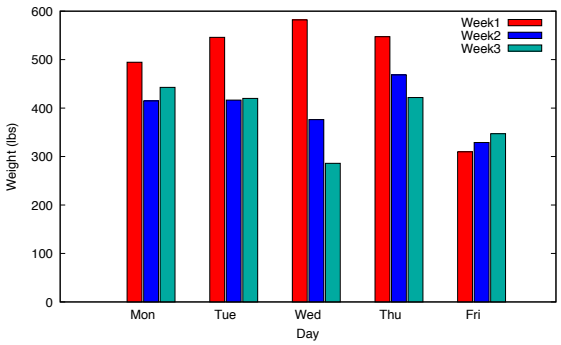
We calculated the mean over each week. To determine if the decrease in waste was significant, we applied a standard Z-Test to the normalized values. Calculation of the Z value is discussed in the original publication [1]. Using a Z-Table and the calculated Z-value for each week when the display was present, we are able to determine the p-value for each week and assess statistical significance. For Week 2 compared to Week 1, the p-value was determined to be .035; the results are significant at the .05 level. For Week 3 compared to Week 1, the p-value is .046; the results are also significant at the .05 level.

4.2.2 Display 2

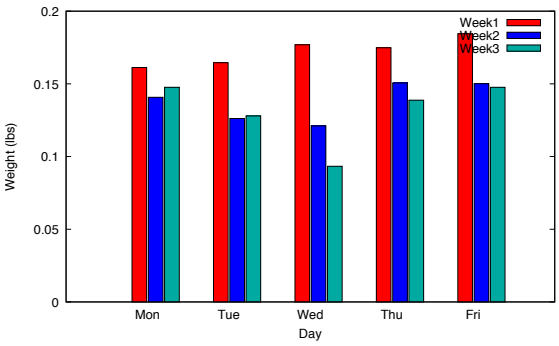
The second display was deployed during Week 5 after a week without the display (Week 4). During Week 4 we collected data but did not include a display; this was done to reduce behavior effects from the previous display from overlapping with the new display. Week 5 introduced the new display.

Behavioral Results Figure 14 compares the waste generated during Week 5 with the waste generated during Weeks 1 and 4. In each graph, the x-axis shows the day of the week in which the data was collected. In Figures 14(a) and 14(c), the y-axis shows the overall amount of food wasted in pounds. In Figures 14(b) and 14(d), the y-axis shows food wasted (in pounds) per meal sold, to normalize the data. Compared to Week 4, when no display was present, customers actually wasted slightly more food–the opposite of the desired result. However, compared to the baseline before the first display was introduced, there was negligible change. Table 4 compares the impact of this display to the initial baseline (Week 1), the second week of the first display (Week 3), and the week just prior to deploying Display 2 (Week 4). None of these proved to be statistically significant for lowering waste.

There are two theories as to why this display was unsuccessful at lowering food waste. The first is the connotation of pizza, versus garbage. Garbage is undesirable; therefore, it is easy for customers to perceive the growing number of garbage bags on a screen as

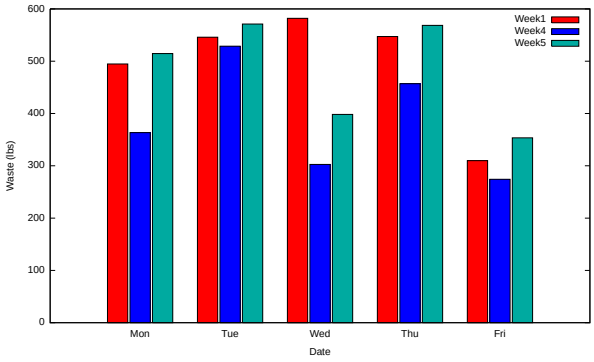


(a) Total Wasted per Day (lbs)

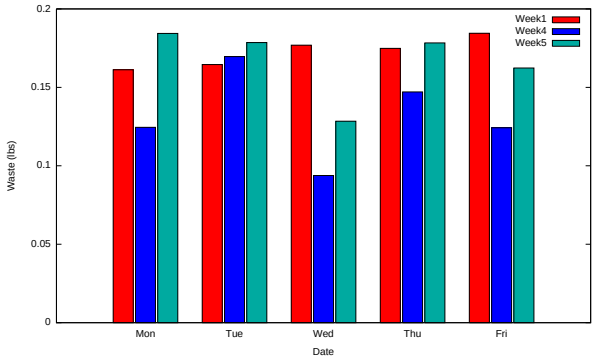


(b) Normalized (waste per meal, lbs)

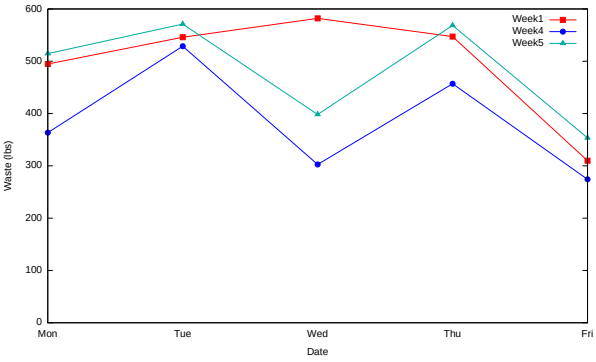
Figure 13: Food Waste With Respect To Display 1



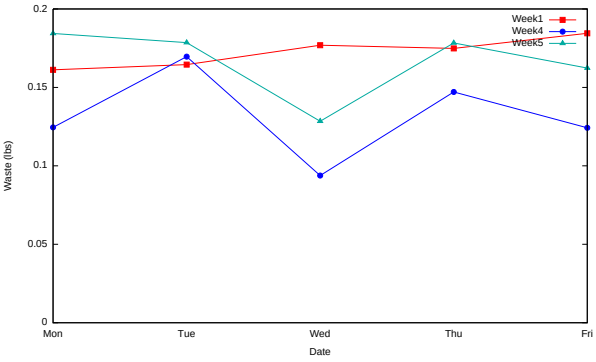
(a) Total Wasted per Day (lbs)



(b) Normalized (waste per meal, lbs)



(c) Total Wasted per Day (lbs)



(d) Normalized (waste per meal, lbs)

Figure 14: Food Waste With Respect To Display 2

Table 4: Comparing Display 2 (Week 5) to Weeks 1,3&4

	Week 1	Week 3	Week 4
Overall Mean (lbs)	0.16939954	0.148733991	0.14913039
Sample Mean (lbs)	0.16640566	0.16640566	0.16640566
Overall StdDev (lbs)	0.01673104	0.028363066	0.03031837
n (days)	5	5	5
Z	-0.40012464	1.393187366	1.27410199
p-value	0.3446	.9177	0.8980

something negative. Pizza, on the other hand, is desirable (especially among college students). The image of pizza multiplying on the display may invoke positive feelings, rather than a sense of alarm. In future work, it would be interesting to test this concept with a less desirable food.

4.2.3 Display 3

The third display for Case Study #1 was evaluated over Weeks 6 and 7, directly following the evaluation of Display 2. During Week 6, there was no display present. As before, this was done to eliminate behavioral effects associated with the previous display from overlapping with the new display. Week 7 introduced the new display. The purpose of this display is to test the "story" component of the SUCCES concepts (discussed in Section 2.2) and humanize the issue.

Behavioral Results Figure 15 compares the waste generated during Weeks 1, 6, and 7. The y-axis in Figures 15(a) and 15(c) shows total waste in pounds, while the y-axis in Figures 15(b) and 15(d) shows waste per meal. For all plots, the x-axis shows the day of the week in which the sample was collected. Week 1 is the baseline week; data was collected this week without a display and before any of the displays were introduced. Week 6 was the week just before this display was introduced; no display was present during this week. As with the previous evaluations, this was done to eliminate lingering effects from the previous display. Week 7 was the week this display was present. In general, the most waste was generated during Week 1, before any of the displays were introduced. An outlier is seen in the data collected on Tuesday of Week 6. This value is much lower due to the technical difficulties experienced that day, not because of an actual reduction in waste. Aside from this outlier, both the total and normalized waste was lowest each day for Week 7.

Table 5 shows the information used to calculate the p-value when comparing the normalized waste each day from Week 7 to the baseline data (Week 1), the first display week (Week 3), and the gap week preceding the display (Week 6). Once again, overall mean is the average for the display week and the comparison week, sample mean is the average for Week 7, overall standard deviation is the standard deviation for the samples across both weeks being compared, and n is the number of days included in the sample. We see that a statistically significant decrease was observed between the baseline week (Week 1, before any displays were introduced) and the week of the display (Week 7), with $p < .05$. There was not a statistically significant improvement from the display shown during Week 3 (Display 1), nor from the week just before the display was implemented (Week 6, no display after the introduction of displays). When comparing Week 6, we calculated the results both with and without the outlier. Neither confirmed $p < .05$; the results listed in the table are those with the outlier excluded.

4.2.4 Significance at Atlantic Dining Hall

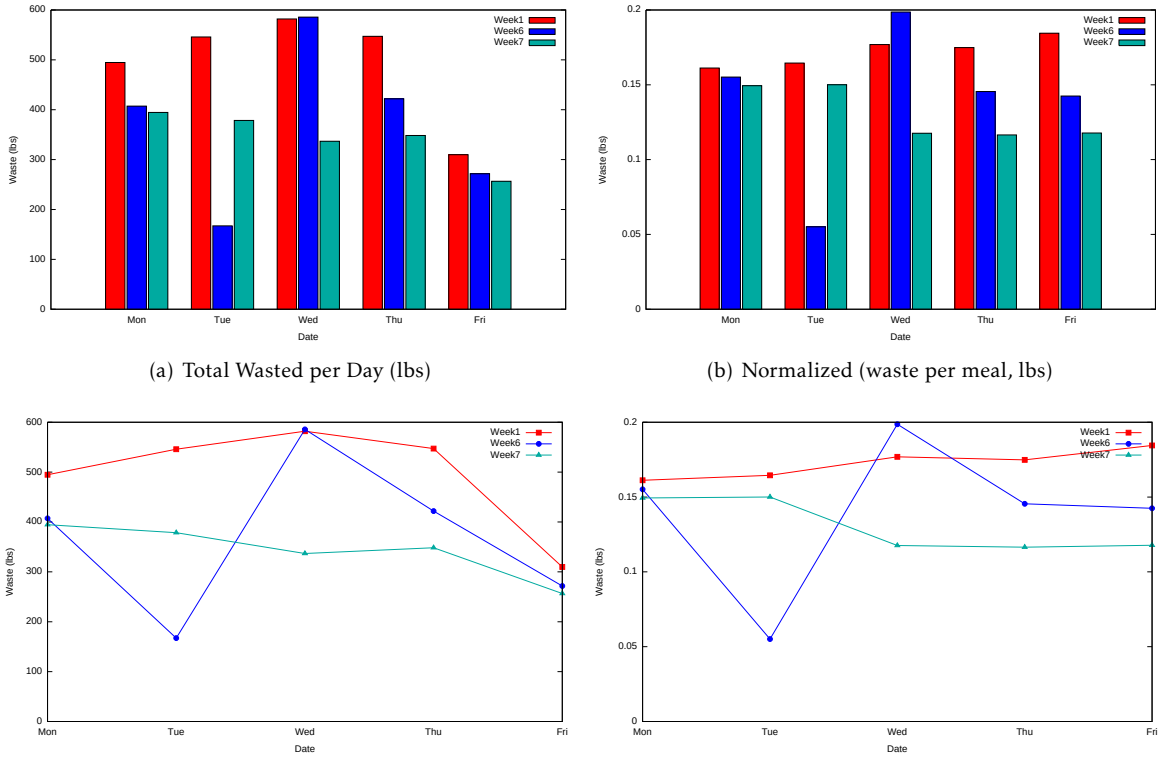
To assess the overall affect of using the displays to affect behavioral change and reduce waste, we compare the two displays that had a significant impact compared to the baseline data. Using three weeks of data, in which two of the weeks used the display, and one week did not, would skew the average of the data toward the waste generated while the displays were present. Therefore, we add to the comparison an additional baseline week from the preliminary deployment in which there was no display, no technical difficulties, and there was a similar packet reception rate. We call this Week 0.

A test of statistical significance was carried out across Week 0 (pre-display), Week 1 (pre-display), Week 3 (display from Case Study #1), and Week 7 (display from Case Study #3). The results of the test are shown in Table 6. Week 0 was too far in the past for the cafeteria to provide us with the exact number of meals sold each day. To normalize the data from that week, we took the average, maximum, and minimum number of customers over the seven weeks data was collected and used those values for normalization. In the table, we show the statistical significance results using the average, maximum, and minimum values for normalization. Assuming the number of customers during that week was near the average, we found that $p = .0174$. If we assume the number of meals sold that week was closer to the maximum observed during the case studies, $p = .0244$. Finally, if we assume the number of meals sold was closer to the minimum observed during the case studies, $p = .0188$. In all three scenarios $p < .05$. Therefore, we can say with confidence at the $p < .05$ level that our displays generated a reduction of waste compared to the baseline.

4.3 Case Study #2

Ideally, WASTE REDUCE would be able to affect behavior in a variety of age groups and settings. We carried out a second case study in the cafeteria at A.D. Henderson University School. The school serves meals for students from kindergarten to eighth grade; however, on most days, only kindergarten through third grade eat in the cafeteria. (Secondary seating is located outside; therefore on rainy days other grades may eat inside.) Once again, we collect one week of data in which the display is not present to establish a baseline, and then collect data for one week with the display present. For this case study, we used Display 2 from Case Study #1 to see if changes in age may alter the effectiveness of the display.

Although this display was unsuccessful in Atlantic Dining Hall, it was believed that of the displays evaluated, the students would relate to this display the most. Display 1 was unlikely to engage young students, who may not have a concept of environmental protection to connect the concept of food waste littering the beach. The storyline for Display 3 was not developed with young children in mind. Although the text and plot



(a) Total Wasted per Day (lbs)

(b) Normalized (waste per meal, lbs)

(c) Total Wasted per Day (lbs)

(d) Normalized (waste per meal, lbs)

Figure 15: Food Waste With Respect To Display 3

Table 5: Comparing Display 3 (Week 7) to Weeks 1,3&6

	Week 1	Week 3	Week 6
Overall Mean (lbs)	0.15233164	0.1316661	0.14480004
Sample Mean (lbs)	0.13226988	0.13226988	0.13226988
Overall StdDev (lbs)	0.026049314	0.02047648	0.02628205
n (days)	5	5	5
Z	-1.722098	0.06593393	-1.0660622
p-value	0.0427	0.5239	0.1446

Table 6: Comparing Waste (Pre-Display vs. Post-Display)

	Average	Maximum	Minimum
Overall Mean (lbs)	0.149672203	0.145258053	0.16064566
Sample Mean (lbs)	0.121939527	0.121939527	0.121939527
Overall StdDev (lbs)	0.032224476	0.028959416	0.04561033
n (days)	5	5	5
Z	-2.108053109	-1.972363293	-2.0787016
p-value	0.0174	.0244	.0188

is fairly simple, it would likely be a challenge for students just learning to read. In addition, the storyline, which follows college or high-school aged students trying to complete a group project, is unlikely to appeal to such a young audience.

4.3.1 Scale Accuracy

Details concerning the accuracy of the scales are presented in Section 3.2.2, and a comparison of the observed versus expected weights of known objects for Case Study #2 is shown in Table 2. Trash Can #1 was found to be perfectly accurate at best, and .055 lbs at worst. Trash Can #2 was found to be accurate within .001 lbs best, and within .032 lbs at worst. This is sufficient to maintain high confidence in the readings from the scale.

4.3.2 Behavioral Significance

Figure 16 illustrates what is typically thrown away during the lunch hour at A.D. Henderson. Unlike Atlantic Dining Hall, there is a higher probability for food packaging, containers, and bottles to be disposed with the food. Also, A.D. Henderson uses disposable trays, which are thrown into the trash can with the food waste. In the figure, it can be observed that the trays, along with packaging containers are included in the waste. Although we were unable to estimate the weight of containers and bottles, we were able to estimate the weight of the trays. We measured the weight of an individual tray and discovered it weighed approximately 0.044 pounds (0.020 kg). Multiplying the number of students who purchased a meal each day by the weight of an individual tray allows us to estimate the total weight of tray waste and exclude it from the analysis.

Figure 16: Observed Waste (A.D. Henderson)

There was no significant change in behavior for the students at A.D. Henderson. Table 7 summarizes the waste accumulated during the case study. The first 3 columns show the total waste, tray waste, and normalized waste, respectively, for the first week of the case study. During this week, no display was present. This is our baseline for comparison. The next 3 columns show the same information for the week the display was present. Since tray waste is unavoidable, this value was subtracted from each day's total waste during the normalization process. Once again, we use the number of students who purchased a meal to estimate the tray waste, and then normalize the total waste to pounds per person. On average, the cafeteria generated 12.50 lbs (5.67 kg) of tray waste per day, and approximately 0.10 lbs (0.05 kg) of waste per person. When the display was present, the students wasted approximately 0.009 lbs more per person. This is not significant.

One thing that may have swayed the results was an incident that occurred during day 5 of the baseline week. On this day, the school was placed on a soft lock-down. During the lock-down, students were escorted to the cafeteria, where they picked up food, and were then escorted back to their classrooms to eat. All waste disposed by these students was deposited outside of the cafeteria. However, the number given for students purchasing meals that day includes both the students who ate in the cafeteria, and those who did not. This means that fewer trays and fewer students contributed to the 21.682 lbs of waste than we estimate. Based on this disturbance, it is likely that there was no change, or an insignificant decrease in behavior for the students.

There are several other differences between the case study at Atlantic Dining Hall and A.D. Henderson that may have affected these results. Everyone who dines at Atlantic Dining Hall eats a meal prepared by the dining hall. However, some of the students at A.D. Henderson bring their lunch from home. This creates three major differences: (i) packed lunches are likely the source of the food packaging containers, (ii) portion sizes for the food brought from home are not likely to be consistent, or to follow recommended guidelines, and (iii) these children are not included in the head count for meals sold.

Another difference is that the children at A.D. Henderson are required to take a full meal. This means they had less control over what they were given, and their ability to waste less. For example, during the case studies, students at Atlantic Dining Hall could view the display, decide they never finish their mashed potatoes, and begin requesting a smaller portion, or no portion at all to aid in reducing waste. Students at A.D. Henderson could not request smaller portions or decline portions of their meal. Therefore, students may have been impacted by the screen, but were powerless to change the outcome.

Age is another factor that may have played a role. Children are generally more inquisitive about changes in their surroundings and were visibly more curious about the changes as the equipment was being installed. However, children are also less likely to understand the significance of lowering food waste.

When food is disposed at Atlantic Dining Hall, cus-

Table 7: Waste at A.D. Henderson University School (in lbs)

	Baseline (No Display)			With Display		
	Total Waste	Tray Waste	Normalized	Total Waste	Tray Waste	Normalized
Day 1	48.401	11.220	0.146	36.258	13.552	0.074
Day 2	30.050	13.068	0.058	48.950	13.552	0.115
Day 3	48.911	12.848	0.124	40.012	12.056	0.146
Day 4	62.777	13.728	0.157	50.586	11.396	0.151
Day 5	27.401	11.748	0.059	33.907	11.704	0.083
Average	43.508	12.522	0.108	41.943	12.452	0.114

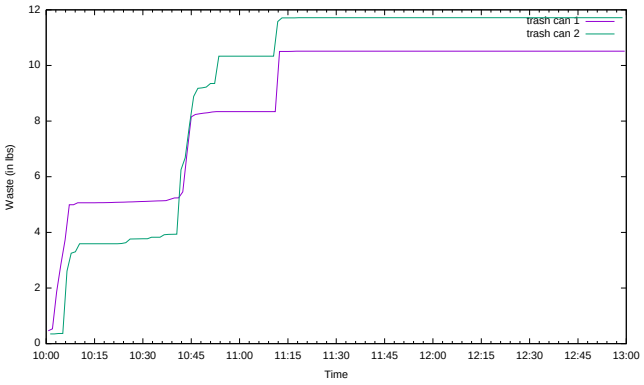
Table 8: Comparison of Food Waste by Grade

Time	Grade	Baseline (lbs)	With Display (lbs)
10:20-10:52	2nd Grade	13.0471711	15.8778712
10:37-11:11	Kindergarten	7.09731827	7.1770412
11:15-11:47	3rd Grade	10.9046479	8.88260616
11:56-12:30	1st Grade	16.4856573	10.0050221

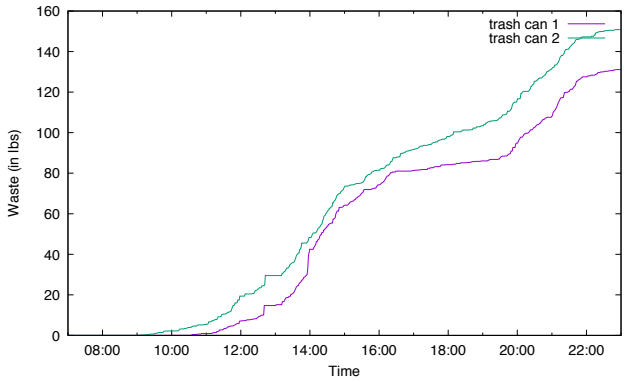
tomers come and go intermittently. Due to the randomized nature of individuals entering and exiting the dining hall, the display is constantly updating, and customers can easily view their individual contribution to total waste. On the other hand, A.D. Henderson has set meal times for the students, and everyone disposes of their waste at the same time. A comparison of the daily waste curve for A.D. Henderson versus Atlantic Dining Hall can be seen in Figure 17. It can be observed that due to the organized nature of the meal times at A.D. Henderson, the graph follows a step pattern; waste increases rapidly as students dispose their waste at the end of their lunch, then flat-lines until the end of the next lunch. Since the display was located at the front of the room, and the trash cans were located at the rear, it is possible students only saw how much the classes before them wasted.

4.3.3 Breakdown By Grade

As noted previously, students at A.D. Henderson eat and dispose of their meals at set times. This allows us to more easily parse out data on food waste by grade. Lunch times and grades are provided in Table 8. Since the head count for meals provided does not include students who bring meals from home and is not broken up by grade level, the data presented is not normalized. Because the data is not normalized, we cannot make inferences across grade levels. For example, the kindergarten lunches produced the least amount of waste; however, they may also have the least number of students attending lunch. These numbers do, however, provide a better view of how the display affected each grade. As expected, there was little to no change for the kindergarten lunch. This age group was likely too young to understand the concept. The students from the second grade lunch appeared to waste about 2 lbs more, on average, after the display was implemented. The first and third grade students saw an average decrease in waste of approximately 2 lbs and 6 lbs, respectively. Although a decrease was viewed for these grades, when tested for statistical significance, it was found that for both groups, $p > .05$.



(a) A.D. Henderson University School



(b) Atlantic Dining Hall

Figure 17: Pattern of Waste Over a Day

5 Conclusion

We have presented WASTE REDUCE, which is a wireless sensor network to provide real-time data on post-consumer food waste. We address software and hardware challenges, such as the communication protocol,

wireless connection stability, and power consumption, to ensure WASTE REDUCE is accurate, robust, and sustainable. We also provide three ambient displays designed to evoke behavior change in customers to reduce food waste. We conducted two case studies to evaluate the efficacy of these displays to generate a change in behavior among cafeteria customers. Analysis of the data before and after the displays were introduced confirms behavioral change was affected for two of the three displays developed for the system.

6 Acknowledgments

This work was supported by the NSF grant CNS-1541917. The authors would like to acknowledge Chancey Kelley and Nicholas Lee Yee for their work in designing the electronic enclosure. The authors would also like to thank Chartwells and A.D. Henderson University School for their assistance and cooperation in using their respective dining halls to carry out each case study.

References

[1] Shiree Hughes, Jiannan Zhai, and Jason O. Hallstrom. Waste auditing sensor technology to enhance the reduction of edible discards in university cafeterias & eateries. In *2018 IEEE International Conference on Smart Computing, SMARTCOMP 2018, Taormina, Sicily, Italy, June 18-20, 2018*, pages 344–349, 2018.

[2] Dana Gunders. Wasted: How america is losing up to 40 percent of its food from farm to fork to landfill. *Natural Resources Defense Council Issue Paper*, August 2012. IP: 12-06-B.

[3] U.S. and World Population Clock. U.S. Department of Commerce, 2016. http://www.census.gov/popclock/.

[4] Lisa R. Young and Marion Nestle. The contribution of expanding portion sizes to the us obesity epidemic. *American Journal of Public Health*, 92(2):246–249, 2002.

[5] UC-Davis. UC Davis profile, March 2016. https://www.ucdavis.edu/sites/default/files/upload/files/uc-davis-student-profile.pdf.

[6] Samantha Lubow, Shaina Forsman, Steganie Scott, Linnea Whitney, Isaac O'Leary, and Amy Wang. Love food, dont waste: Resident dining commons food waste report, 2015. http://dining.ucdavis.edu/documents/SP15WasteAuditReport.pdf.

[7] UC-Davis. Waste reduction and elimination, 2015. http://dining.ucdavis.edu/sus-recycling.html.

[8] GT. Admissions & enrollment, 2016. http://factbook.gatech.edu/quick-facts/admissions-enrollment/.

[9] GT. Campus services fall 2012 semester report, January 2013. http://campusservices.gatech.edu/publications/Documents/fallsemesterreport0113final.pdf.

[10] LeanPath. What is LeanPath?, 2015. http://www.leanpath.com/product-video/.

[11] MSU. Msu facts, 2016. https://msu.edu/about/thisismsu/facts.html.

[12] Simon Anderson, Alex Parker, Lisa Schindler, Lindsay Stolkey, and Kaite Ward. MSU: Wasted students? or wasted food. *Michigan State University*.

[13] Carla Iansiti, Allison Stawara, and Robbia Pipper. Food waste impacts msu, 2013. *National Association of College & University Food Services 2013 Regional Conference*. https://eatatstate.com/sites/default/files/pdf/2013%20NACUFS%20Regional%20Conference.pdf.

[14] FIU. Sustainability solutions opportunity assessment report. *Waste Management Sustainability Services*, May 2011. Florida International University. http://gogreen.fiu.edu/_assets/documents/topics/Waste-Audit-Report.pdf.

[15] UM. Student enrollment fall 2015, 2016. http://www.miami.edu/index.php/about_us/fast_facts/student_enrollment_--_fall_2015/.

[16] UM. Sustainability interim report, 2014. Office of Sustainability. https://umshare.miami.edu/web/wda/greenu/DOCUMENTS/Sustainability%20Interim%20Report%202013-2014_UMWeb.pdf.

[17] UM. Green waste - food waste, 2016. Office of Sustainability. http://www.miami.edu/finance/index.php/green_u/recycling/green_waste_-_food_waste/.

[18] Oxford English Dictionary. habitutation, n 1.1. https://en.oxforddictionaries.com/definition/habituation, 2016.

[19] Kevin Vaughan and Nicholas Tarrier. The use of image habituation training with post-traumatic stress disorders. *The British Journal of Psychiatry*, 161(5):658–664, 1992.

[20] Han X Lin, Yee-Yin Choong, and Gavriel Salvendy. A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour & information technology*, 16(4-5):267–277, 1997.

[21] Chip Heath and Dan Heath. *Made to Stick: Why Some Ideas Survive and Others Die*. Random House, 2008.

[22] John Stasko, Todd Miller, Zachary Pousman, Christopher Plaue, and Osman Ullah. Personalized peripheral information awareness through information art. In *International Conference on Ubiquitous Computing*, pages 18–35. Springer, 2004.

[23] Mark Weiser and John Seely Brown. The coming age of calm technology. In *Beyond calculation*, pages 75–85. Springer, 1997.

[24] Dan Ariely. Controlling the information flow: Effects on consumers' decision making and preferences. *Journal of Consumer Research*, 27(2):233, 2000.

[25] Franck L Lewis et al. Wireless sensor networks. *Smart environments: technologies, protocols, and applications*, pages 11–46, 2004.

[26] Ian F Akyildiz and Mehmet Can Vuran. *Wireless sensor networks*, volume 4. John Wiley & Sons, 2010.

[27] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

[28] Raspberry Pi Foundation. Raspberry pi. https://www.raspberrypi.org/, 2017.

[29] Massimo Banzi, David Cuartielles, Tom Igoe, Gianluca Martino, David Mellis, et al. Arduino. https://www.arduino.cc/, 2005.

[30] Beagleboard.org. Beaglebone black. https://beagleboard.org/black, 2016.

[31] Niels Reijers, Gertjan Halkes, and Koen Langendoen. Link layer measurements in sensor networks. In *Mobile Ad-hoc and Sensor Systems, 2004 IEEE International Conference on*, pages 224–234. IEEE, 2004.

[32] Christopher Plaue, Todd Miller, and John Stasko. Is a picture worth a thousand words?: An evaluation of information awareness displays. In *Proceedings of Graphics Interface 2004*, GI '04, pages 117–126, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

[33] Andrew Dahley, Craig Wisneski, and Hiroshi Ishii. Water lamp and pinwheels: Ambient projection of digital information into architectural space. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, CHI '98, pages 269–270, New York, NY, USA, 1998. ACM.

[34] Jennifer Mankoff, Anind K. Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 169–176, New York, NY, USA, 2003. ACM.

[35] Jeremy M Heiner, Scott E Hudson, and Kenichiro Tanaka. The information percolator: ambient information display in a decorative object. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, pages 141–148. ACM, 1999.

[36] Hannah Pileggi, Charles D Stolper, J Michael Boyle, and John T Stasko. Snapshot: Visualization to propel ice hockey analytics. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2819–2828, 2012.

[37] HT Sensor Technology Co. Limited. Tal201 parallel beam load cell. http://www.htc-sensor.com/products/87.html, 2016.

[38] AVIA Semiconductor (Xiamen) Co. LTD. Hx711: 24-bit analog-to-digital converter (adc) for weigh scales. http://www.aviaic.com/UpLoadFile/hx711_brief_en.pdf, 2016.

[39] Digi International Inc. Xbee-pro 900hp. https://www.digi.com/products/xbee-rf-solutions/modules/xbee-pro-900hp, 2016.

[40] Atmel Corporation. Sam4l xplained pro evaluation kit. http://www.atmel.com/tools/ATSAM4L-XPRO.aspx, 2016.

[41] TP-Link Technologies Co. Ltd. 150mbps high gain wireless usb adapter: Tl-wn722n. http://www.tp-link.com/en/products/details/cat-11_TL-WN722N.html, 2016.

[42] Inc. MongoDB. Mongodb. https://www.mongodb.com/, 2016.

[43] Author One, Author Two, and Author Three. Title of paper. In *Conference*, page Pages, Year.

[44] Mike Bostock. Data-driven documents. https://d3js.org/, 2015.

[45] Beltsville Human Nutrition Research Center. Usda food composition databases. *United States Department of Agriculture*. https://ndb.nal.usda.gov/ndb/, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 65-69 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A Relation Extraction System for Indian Languages

Ajees Arimbassery Pareed*, Sumam Mary Idicula

Department of Computer Science, CUSAT, 682022, India

ARTICLEINFO

Article history:
Received: 04 January, 2019
Accepted: 28 February, 2019
Online: 15 March, 2019

Keywords:
Convolutional Neural Networks
Natural Language Processing
Word embedding
Relation extraction.

ABSTRACT

Relation Extraction is an important subtask of Information Extraction that involves extracting significant facts from natural language text. Extracting structured information from the plaintext is the ultimate goal of IE systems. The Indian language content on the internet is increasing day to day. Extracting relevant information from this huge unstructured data is a challenging task especially when the business firms are interested in ascertaining public view on their products and processes. The primary objective of relation extraction systems is to find those entities which can be targeted through social networking and digital marketing. Cannibalisation of the product is nowadays done using these Social Networks. Different methods are proposed and experimented for Relation extraction problems. In this paper, we propose a Relation Extraction system using Convolutional Neural Networks. Deep learning based methods have produced state of the art results in many domains. Training and testing are conducted using the shared corpus provided by 'ARNEKT-IECSIL 2018' competition organisers. The evaluation results show that the proposed system could outperform most of the reported methods in the competition.

1 Introduction

Internet is the fastest growing resource on the planet. Lots of information are added to the web every second. However, this information is stored in an unstructured manner. Retrieving the relevant information from this unstructured text is a challenging task that invites the focus of Language researchers. Information extraction, a branch of Artificial Intelligence deals with this challenge [1]. IE transforms the unstructured text into a structured form that can be easily handled by machines. Relation Extraction is one of the subdomains of IE. It is the process of identifying the relation between two entities in a document. There are two major types of RE namely-closed domain and open domain RE. Closed domain RE considers only a closed set of relationships between two arguments while the open domain RE systems use an arbitrary phrase to specify a relationship [2].

Relation Extraction is one of the subtasks of Information Extraction. The occurrence of entities in a sentence is always through well-defined relationships. Automatic identification of such relationships is what we call as the task of relation extraction. Relation Extraction also helps in getting the structured information from the unstructured text. It is very similar to Information Extraction with the exception that IE additionally requires the removal of repeated relations. Applications of Relation Extraction systems include construction of knowledge bases, Question answering systems, text summarisation, etc. Construction of knowledge bases is a laborious, time-consuming project that demands domain expertise. Automatic extraction of relationships and concepts from text documents helps to reduce the time and domain expertise needed for the task. Question answering systems also make use of relation extraction systems since relations can provide a clue about the answers to most of the questions. Similarly, relation extraction systems can be employed in areas like textual entailment, gene-disease prediction, protein-protein interaction, etc.

Text classification is the primary area of research where the supervised machine learning algorithms are explored in Natural Language Processing. It is a very active research area both in industry and academia. Examples of such classification tasks include sentiment analysis from social media text, detection of spam emails, categorisation of customer queries, auto tag-

*Corresponding Author: Department of Computer Science, CUSAT, 9061859697 & ajeesap87@gmail.com

https://dx.doi.org/10.25046/aj040208

ging of news articles, etc. In this work, we have tried to move towards a direction which is not much explored in the case of Indian languages.

The structure of this article is as follows. Section 2 briefly reviews the related works and details about the dataset. Section 3 explains the proposed method and section 4 illustrates the experiments and results. Finally, section 5 concludes the article along with some routes for future works.

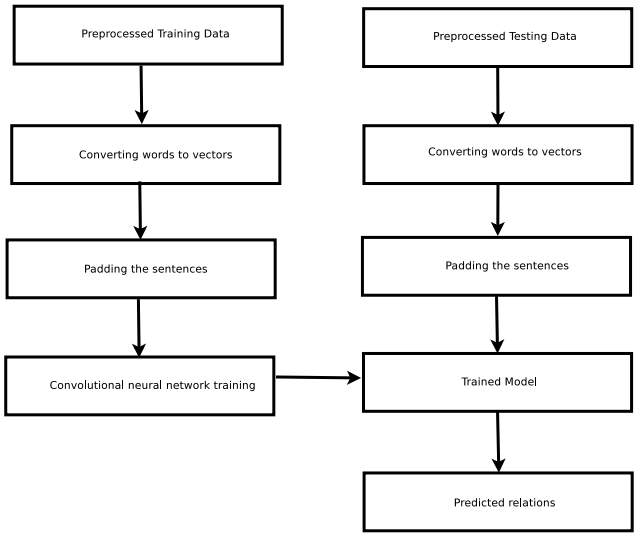


Figure 1: Architecture of the proposed system.

2 Related Works

Various studies are conducted to extract the semantic relations between entities in a text document. Starting from the straightforward rule-based approach to more complex supervised approaches and also semi-supervised approaches are explored in the literature [3]. Rule-based systems employ a naive approach towards the relation extraction task. They construct a set of possible rules for relation extraction by looking at some examples. Hearst, who applied this approach to find the hyponyms of words could achieve an accuracy of 66% on his task [4]. The major issue associated with this technique is that it needs a lot of rules to accomplish the task. We have to continuously update the rule set whenever we find an exception to the existing rules. And more notably, we have to redo this entire work for the other kinds of relations.

On the other hand, supervised relation extraction systems build the model with the help of already tagged corpus. Set of features are designed for each sample in the training data with the help of domain experts. This feature set act as input to the learning algorithms. The key idea behind the supervised learning algorithms is to model the relation extraction task as a classification problem and train the classifier with different algorithms (SVM, Naive Bayes, etc.) available for supervised learning. These classifiers can be trained using a different set of features selected after performing a textual analysis of the labeled data. Different features considered for the study include lexical features, syntactic features, dependency features, entity features, etc [3].

Supervised methods suffer from the problem of availability of enough labelled data. If we do not have enough labelled data to train our classifier, results will be poor. The solution to this issue is the bootstrapping technique. In this technique, we will start with some seed instances of training data, which is manually tagged data used for the first phase of training. We train our classifier with seed instances and learn the classifier. This classifier is used to test more unlabelled data, and get more train examples by adding the test results to the training set. Thus, the training set will expand up to a sufficient amount. This approach is called a semi-supervised learning. 'DIPRE' is an example of the semi-supervised system employed for the relation extraction task [5]. It tries to extract author-book relationship from web text. However, the current trend in relation extraction is based on reinforcement learning. Xiangrong Zeng et al.[6] reports an improvement of 13.36% over the baseline models with the help of reinforcement learning.

The shared task is divided into two subparts say task-A and task-B [7]. Task-A deals with the identification of named entities from the raw text and task-B deals with extracting relation amongst the entities in a sentence. Both these tasks come under the domain of Information Extraction (IE), which is an area under constant research. The growth of research in this area leads to the advancement of applications like information search, question answering, document summarization, etc. Five Indian languages are considered for this shared task. They are Tamil, Hindi, Kannada, Telugu, and Malayalam. It is well known that IE works significantly well with languages like English from applications like Google search, frameworks like Stanford CoreNLP, OpenNLP and many more. The same does not hold good for Indian Languages due to its morphologically rich nature and agglutinative structure. Hence, the objective of this shared task is to improve the Information Extraction systems for Indian languages.

The shared dataset contains data from five different Indian languages [8]. The training data for task-B is a set of files in plain text format. Each file consists of sentences and their corresponding labels. Each language has more than 25000 samples of training data. Statistics of the training data for the task-B is shown in figure 2. The testing data contains two files say test1 and test2. Test1 is for pre-evaluation, and test2 is for final evaluation. The statistics of the test data is given in table 1.

3 Proposed Method

The problem is framed as a sentence classification problem with classes as relations. The number of classes is equal to the number of relations in the tagged data. Convolutional neural networks are used to build the model. They are a class of neural networks that have

Table 1: Data statistics

Language	Train	Pre-eval	Final-eval
Hindi	56775	18925	18926
Kannada	6637	2213	2213
Malayalam	28287	9429	9429
Telugu	37039	12347	12347
Tamil	64833	21611	21612

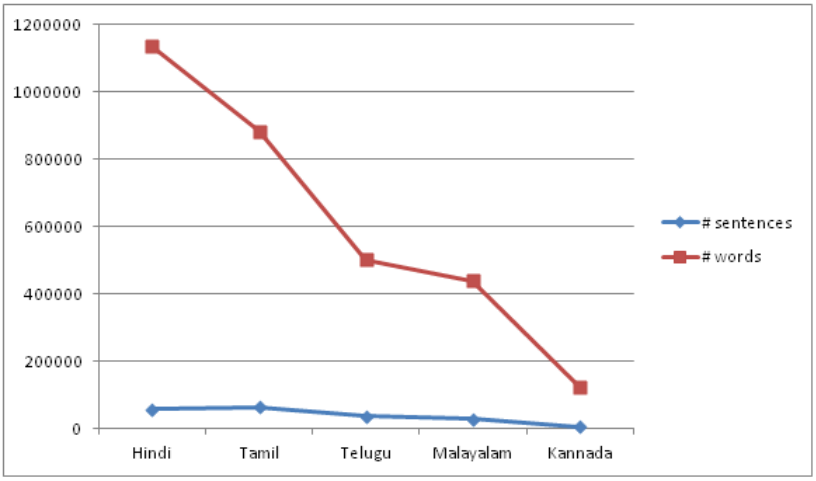


Figure 2: Statistics of the training data

proven very effective in areas such as image classification and pattern recognition. Figure 3 depicts the easiness of deep learning based text processing systems over the traditional machine learning approaches. They can capture the local texture within the text and can be used to find the representative patterns in a text document [9]. The most important property of CNN is preserving the 2D spatial orientation in computer vision problems. But these orientations have a one-dimensional structure in the case of texts.

The architecture of the proposed method is shown in figure 1. The first phase is the preprocessing phase, where words in each sentence are separated using NLTK word splitter. Since words are symbolic units, it can't be directly fed to neural networks. Hence words are converted into numeric values(vectors) using word2vec [10]. The sequences of vectors are then padded with zeros to make it of uniform length. Finally, the padded sentences and their corresponding labels are provided to CNN for training. After training, the model file is saved for testing. In the testing phase, the test data is passed through operations similar to that of the training data and are provided to the saved model for prediction. The model then predicts the label for each sentence indicating the relation within the sentences.

4 Experiments and Results

The preprocessing stage contains operation like trimming, stopword removal, etc. Trimming is the process of removing unwanted symbols from the text. And the most common words are removed in the stop word removal stage. Inputting sequences of raw human alike words will make no sense to computers. For that reason, the raw words are converted into vectors of numeric value using word2vec [11]. Word2vec model is built using a manually created corpus of 27 lakhs words. Skip-gram configuration of Word2vec is used to build the model. The model is constructed with a context window size of 10. The dimensionality of the word embedding is fixed as 100 to cope up with the processing power of the machine.

Keras sequential model is used to construct the classifier [12]. The configuration of the constructed convolutional neural network is shown in table 2. The network is designed with four convolutional layers, two max-pooling layers, and two dense layers. The first layer is a convolutional layer for its ability to capture the local context. The following layers are alternate max-pooling and convolutional layers for acquiring the hidden patterns within the sentence. We have used 'Relu' as the activation function to bring nonlinearity. The number of filters used in the first two convolutional layers is 756. And the kernel size is fixed at 7 for the first two convolutional layers and 3 for the remaining layers. The final dense layer is associated with softmax activation units. During the training phase, filters slide over full rows of the word embeddings. CNN automatically learns the values of its filters based on the labels on the training samples.

In our experiments, we have selected the first 90% of the sentences as training data, and the remaining is selected as the testing data. The batch size is fixed at 100. Categorical cross entropy is used as the loss function. We used Adam, the efficient gradient descent algorithm as the optimizer. Dropout is used to

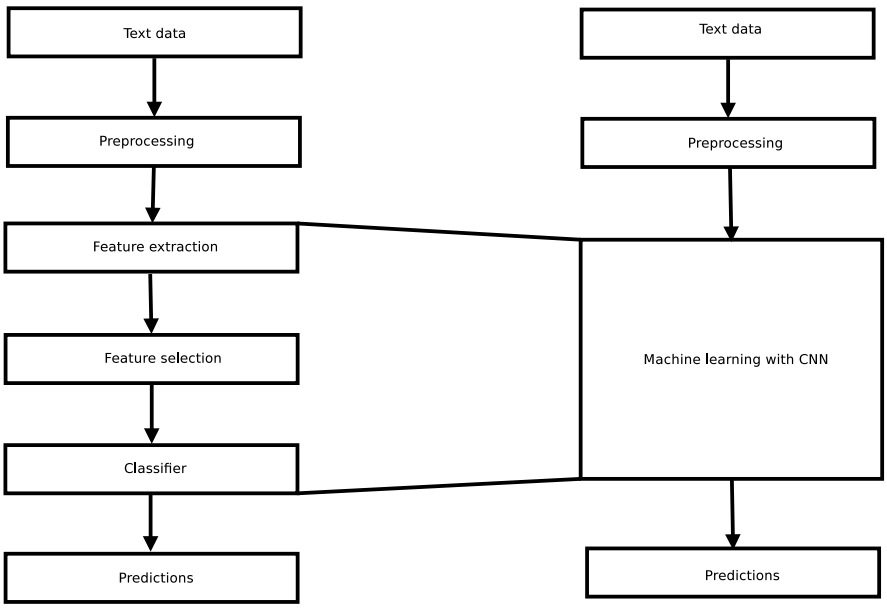


Figure 3: Comparative architecture of deep learning based and non-deep learning based systems

Table 2: Configuration of the CNN architecture

Layers	Output shape	Configuration
conv1d_1	30×756	$7 \times 1, strides 1$
max_pooling1d_1	15×756	$3 \times 1, strides 1$
conv1d_2	15×756	$7 \times 1, strides 1$
max_pooling1d_2	7×756	$3 \times 1, strides 1$
conv1d_3	7×256	$3 \times 1, strides 1$
conv1d_4	7×256	$3 \times 1, strides 1$
flatten_1	1792	–
dense_1	128	Relu activation
dropout_2	128	0.5
dense_2	14	Softmax activation

prevent overfitting [13]. Model is compiled using Tensorflow in the backend. The network is trained for five epochs, and the model file is saved for the testing. Due to the lack of pre-trained word embeddings, we could not complete our work on languages other than Malayalam using pre-trained word embeddings. For Malayalam, we were able to simulate word vectors using our corpus. Moreover, the publicly available pre-trained word embeddings were out of the scope of our machine memory. For languages other than Malayalam, we used an additional embedding layer in front of the convolutional layers.

The proposed system is tested with two test datasets(pre-evaluation and final evaluation). Our system predicts the relation in each input sentence. Table 3 demonstrates the results of our system on both the datasets. Since the final evaluation phase is not available for real-time evaluation, we were not able to get the results for test 2 dataset for languages other than Malayalam. It is clear from the results that our system performance is promising as compared with the performance of other methods reported in the competition. The performance of different relation extraction systems reported in the competition (on test 1) is shown in figure 4. Our system records an average accuracy of 85.62% on test-1 data. From the experimental results, it is obvious that the performance of the system increases with the increase in the training data size. This point is evident from the results of Kannada and other languages. Kannada, which contains the least number of training samples, records the least performance as compared with other languages in the dataset.

5 Conclusion

In this paper, we have discussed a CNN based Relation Extraction system for Indian languages. The exclusive feature of our technique is the use of CNN for relation extraction in Indian languages. The main reason we preferred CNN rather than other traditional feature-based methods is their ability to capture the relations within the sentences. Since deep learning methods require a sufficient amount of training data, the performance of the system can still be improved by increasing the training data size. The performance of the system can also be improved by incorporating word embedding based cluster features into the word vectors. Due to the lack of enough computational power, we could not accomplish that task. Apart from Rela-

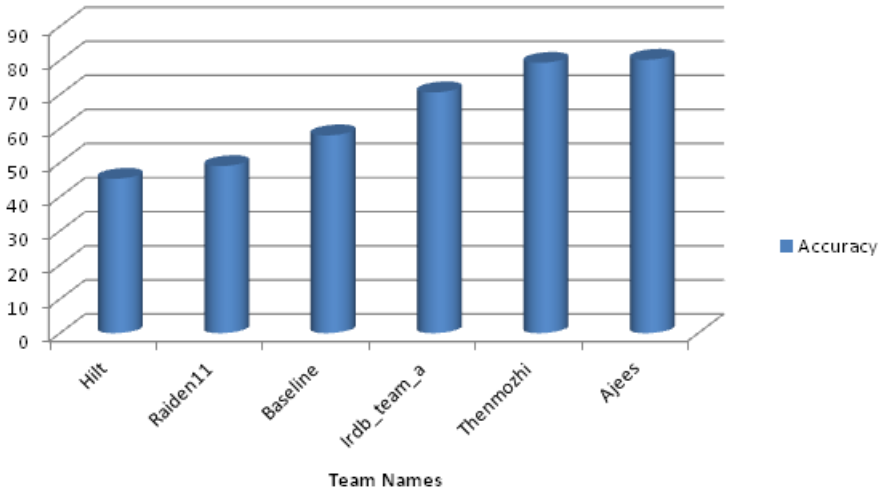


Figure 4: Performance of different relation extraction systems on the shared corpus

Table 3: Results

Test data	Hindi	Kannada	Malayalam	Tamil	Telugu	Average
Test 1 (Accuracy %)	94.72	50.64	80.46	85.76	84.76	85.62
Test 2 (Accuracy %)	NA	NA	77.77	NA	NA	15.55

tion Extraction, Convolutional Neural Networks based deep learning methods can also be applied to various NLP applications like Text classification, sentiment analysis, document labelling, etc.

References

[1] Nita Patil, Ajay S Patil, and BV Pawar. Survey of named entity recognition systems with respect to indian and foreign languages. *International Journal of Computer Applications*, 134(16), 2016.

[2] Manaal Faruqui and Shankar Kumar. Multilingual open relation extraction using cross-lingual projection. *arXiv preprint arXiv:1503.06450*, 2015.

[3] Kush Goyal and Pushpak Bhattacharyya. Literature survey on relation extraction and relational learning.

[4] Barbara Rosario and Marti Hearst. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.

[5] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 129–136. Association for Computational Linguistics, 2006.

[6] Zeng Xiangrong, Liu Kang, He Shizhu, Zhao Jun, et al. Large scaled relation extraction with reinforcement learning. 2018.

[7] HB Barathi Ganesh, KP Soman, U Reshma, Kale Mandar, Mankame Prachi, Kulkarni Gouri, and Kale Anitha. Overview of arnekt iecsil at fire-2018 track on information extraction for conversational systems in indian languages. In *FIRE (Working Notes)*, 2018.

[8] HB Barathi Ganesh, KP Soman, U Reshma, Kale Mandar, Mankame Prachi, Kulkarni Gouri, and Kale Anitha. Information extraction for conversational systems in indian languages - arnekt iecsil. In *Forum for Information Retrieval Evaluation*, 2018.

[9] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[11] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[12] François Chollet et al. Keras. 2015.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 412-424 (2018)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Two Degree-of-Freedom Vibration Control of a 3D, 2 Link Flexible Manipulator

Waweru Njeri*, Minoru Sasaki, Kojiro Matsushita

Mechanical Engineering, Mechanical Engineering, 1-1 Yanagido, Gifu, Japan

ARTICLEINFO

Article history:
Received: 24 October, 2018
Accepted: 2 December, 2018
Online: 15 December, 2018

Keywords:
Inverse control
Strain feedback
System stability
Two-degree-of freedom

ABSTRACT

Considering link vibrations, the main limitation affecting flexible manipulators, this article seeks to make a contribution by presenting an enhanced two degree of freedom vibration controller. This controller uses a filtered right inverse controller in the feedforward and strain feedback controller in the feedback path. The Filtered inverse controller damps transient vibrations while preserving joint trajectories. On the other hand, strain feedback controller ensures a rapid decay of residue vibrations. Modeling of the manipulator was carried out in Maplesim, linearized and inverted in Matlab. Experiments were conducted in the dSPACE environment. Both the simulations and the experimental results showed that the two-degree-of-freedom controller yielded a superior performance over the two controllers individually.

1 Introduction

This journal article, an extension of our original work presented in the 2018 IEEE International conference on Applied System invention(ICASI) [1] seek to make a contribution by presenting a two degree of freedom controller comprising of a filtered inverse controller in the forward path and a Direct Strain Feedback controller(DSFB) in the feedback path. The strength of the proposed methods lies in the fact that it can suppress both the motion induced vibrations, as well as residue vibrations, and it is superior than the individual controllers separately.

Since the dawn of the industrial revolution, over and above making the work easier, man has been thinking of how to replace human labour altogether. This has brought to fruition many kinds of robots and manipulators to fulfil this dream. These robots came handy in carrying out repetitive chores, in hazardous work environment and in heavy industries to mention just but a few. Initially, robots comprised of bulky links driven by huge motors. Thus, they suffered from link inertia or were rather driven at low speeds and were expensive to operate considering the amount of power they consumed and the cost of maintenance. Challenges of rigid robots, except for heavy tasks and for application where low speeds is not a problem, were addressed by the introduction of flexible manipulators.

These Flexible manipulators come with lots of merits over rigid manipulators; like being light in weight hence, small actuators can be used. Also, the maintenance and running costs are low. However, their links are flexible, hence they tend to vibrate especially when operated at high speeds. These vibrations increases with both loading and additional links, due to coupling, which adversely affects the accuracy and duplicity of tasks. Also, link vibration leads to time wastage, hence having to wait for the residue vibrations to decay to healthy levels before the end-effector can be put to use. Link vibrations also raises a safety issue, as continued vibration can lead to mechanical failure due to fatigue, thereby reducing the life span of the manipulator, and posing a risk to the operators.

In a bid to reap all the benefits of the flexible manipulators, a lot research has been done to mitigate link vibrations and all its shortcoming. Researchers [2, 3] introduced a vibration control method for the flexible manipulators based on internal resonance. In their scheme, they designed an absorber comprising of a servomotor, sliding along the links of the manipulator and driving a branched link. At internal resonance, the energy stored in the links in form of vibration energy, is propagated back and forth between the different modes and is dissipated to the absorber. One challenge associated with their method includes the complication in establishing the internal resonance considering the

*Corresponding Author, Waweru Njeri, Mechanical Engineering Department, Gifu University, 1-1 Yanagido, Gifu, Japan
Email: v3812104@edu.gifu-u.ac.jp

https://dx.doi.org/10.25046/aj030649

different loading and trajectories. Another limitation is the feasibility of using this scheme with additional links, and the inertia introduced by the absorber.

In [4], the author proposed a novel filtered input shaping technique, implemented inside the position control loop of a single link flexible manipulator. The controller was developed based on the dominant vibration modes of the manipulator to mitigate the motion-induced vibrations from the high speed operations, which arose also from the flexible nature of the link. The controller was based on a lowpass and a bandpass elliptic filters to attenuate dominant modes in the input signal, and avoid exciting the manipulator at its natural frequencies. With fixed filters, however, the performance of their system would deteriorate with additional loading, since dominant modes are bound to change. The limitations of the fixed filters are addressed in [5]. There exist similar work involving input shaping for vibration control, for example [6–8].

In [9, 10], the authors noted that vibrations occur after the manipulator was brought to a sudden stop. They found that the severity of the induced vibration was largely dependent on the initial speed, speed prior to the stop and the duration of the motion. The higher the speed, the longer the vibrations will ensue, before the end pointer settles to the desired position. In their work, they developed trapezoidal and triangular velocity profiles to reduce joint velocities at the beginning of the operation and prior to stopping. In other words, the trajectory was portioned into acceleration period, constant speed period and deceleration period. The scheme resulted in minimal vibrations naturally as the manipulator decelerated to a stop.

Another popular active vibration control technique is the use of smart structures in form of piezoelectric actuators which are bonded to the root of the manipulator [11]. Authors in [12] made use of shunted piezoelectric transducers to mitigate vibrations. In this type of solution, the piezoelectric transducers are bonded onto the links of the manipulator. Link flexure is converted to corresponding voltages which are used as feedback. Under this broad scheme, alternative vibration suppression methods include: Position positive feedback [13], strain rate feedback control [14], quantitative feedback theory [15], and resonant control [16]. Current trends include using Neural Networks to tune PID gains [17], filtered inverse control [18], robust control, in particular H_∞ together with piezoelectric actuators [19], boundary control [20].

The rest of the article is organized as follows; Modelling of the two link 3D flexible manipulator is introduced in section 2. Development of the inverse model is highlighted in section 3. Application of the developed inverse is applied to the manipulator 4. Simulation and experimental results are presented and discussed in section 6 followed by conclusion in section 7.

2 Modeling and validation of the manipulator

The basic prerequisite for model based controller design is an accurate model of the plant to be controlled. This is an easy feat for simple systems which can easily be obtained from first principles by paper and pen observing the governing equations. However, for complex system especially if they are nonlinear and dynamic is not an easy task.

Advancements in the processing power of modern computers has brought forth another method of obtaining models of complex system by employing symbolic softwares. Examples of such softwares include but not limited to Maple/Maplesim©, Mathematica©, Matlab/Simscape© all based on Modelica© library. In symbolic modeling, mathematical models of common engineering parts such as joints, links, motors with possibilities of customization are dragged onto a workspace to form a complex system. Simulations are carried out considering a practical environment. The strength of this technique lies in its accuracy, simplicity and the possibility of including aspects which cannot be represented mathematically.

The plant reported in this article is a two link, 3D flexible manipulator, with technical description as tabulated in Appendix A.1. It has three rigid joints, driven by dc servomotors and fitted with harmonic drives to reduce joint velocities by a factor of one hundred. It has two flexible links with a variable load attached at the far end of the link number 2.

The control system includes a computer running the Matlab/Simulink and interfaced with dSPACE DSP board, which serves as Servomotors driver via DA converters and collection of data via AD converters. Operation of the system is carried out from the dSPACE control desk environment. The measurement of the joint angles and the velocities were achieved using encoders fitted at the bottom part of the servomotors. Strain data was obtained from strain gauges attached at the root of the links. Strain information was conditioned using wheatstone bridge, filtered and amplified before being transmitted to the computer via AD converters. The control system is configured as in Figure 1.

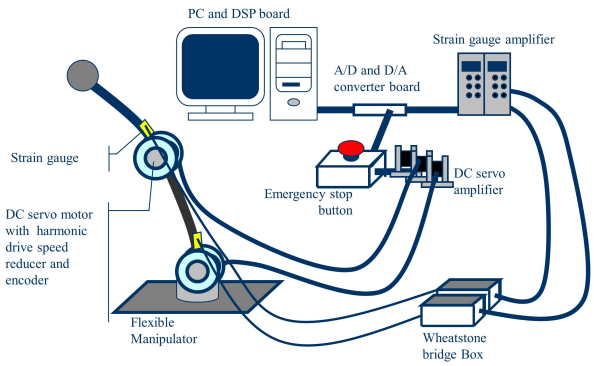


Figure 1: Control system setup

The manipulator and the control system were modelled and linearized in Maple/Maplesim and its

inverse model was developed in Matlab. Validation of the model against the actual manipulator was performed and a perfect agreement was observed between the nonlinear model, linearized model and the actual manipulator, as will be seen later in the results. State space matrices of the linearized model are posted in the Appendix A.3.

3 Development of the inverse system

To develop an inverse model, consider an Linear Time Invariant(LTI) continuous time square system $\sum(t)$, and let the triplet A,B and C be a minimal state-space representation. It is assumed that the system is stable or stabilized by negative feedback.

$$\dot{x}(t) = Ax(t)+Bu(t),\ x(0)=x_0,\ t\in\mathbb{R}^+ \quad (1)$$

$$y(t) = Cx(t) \quad (2)$$

where $x(t)\in\mathbb{R}^n$, $u(t)\in\mathbb{R}^p$, $y(t)\triangleq(y_1,y_2,\cdots,y_p)^T\in\mathbb{R}^p$, $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times p}$ and $C\in\mathbb{R}^{p\times n}$.

Definition 1. Given $\sum(t)$, an LTI system defined above in equations 1 and 2, inversion involves the development of a model $\sum^{-1}(t)$ that yields the input control law $u_f(t)$ to reproduces $y(t)$ when used as the input to $\sum(t)$.

Definition 2. If C_i denotes the i_{th} row of the output matrix C, then the system is said to have a relative degree $r\triangleq(r_1,r_2\cdots,r_p)^T$ if $C_iA^lB=0$, $\forall l<r_i-1$; $1\le i\le p$ [21]. Further, if this holds true in the entire domain in the states, then we say the system has a well defined relative degree.

Following Definition 2 above and assuming that the system has a well-defined relative degree $r=(r_1,r_2,\cdots,r_p)^T$, differentiating the i_{th} output r_i times *w.r.t* time yields

$$y^{(r_i)}=C_iA^{(r_i)}x+C_iA^{(r_i-1)}Bu$$

where C_i is the i_{th} row of the output matrix C for $1\le i\le p$ and the subscripts represent the Lagrange's notation of the r_{ith} derivative in time. Repeating this for all the rows and having the resulting expressions in vector form, we have

$$y^{(r)}=A_xx(t)+B_yu(t) \quad (3)$$

where

$$y^{(r)} \triangleq \begin{bmatrix} y_1^{(r_1)}(t)\\ y_2^{(r_2)}(t)\\ \vdots\\ y_p^{(r_p)}(t)\end{bmatrix}$$

$$A_x \triangleq \begin{bmatrix} C_1A^{(r_1)}\\ C_2A^{(r_2)}\\ \vdots\\ C_pA^{(r_p)}\end{bmatrix}$$

$$B_y \triangleq \begin{bmatrix} C_1A^{(r_1-1)}B\\ C_2A^{(r_2-1)}B\\ \vdots\\ C_pA^{(r_p-1)}B\end{bmatrix}$$

From equation (3), and the fact that B_y is invertible because of the well defined relative degree assumption, the control law is

$$u(t)=B_y^{-1}[y_d^{(r)}-A_xx(t)] \quad \forall\ t\in(-\infty,\infty) \quad (4)$$

There exist a state transformation $T:R^n\to R^n$

$$x(t)=T\begin{bmatrix}\zeta(t)\\ \sigma(t)\end{bmatrix}^T$$

which decomposes the states into internal dynamics(system states, which are not directly controlled by the input u(t)), $\sigma(t)$ and the external dynamics $\zeta(t)$, (i.e, the output and its derivatives in time up to (r_i-1)) as

$$\zeta=[y_1,\dot{y}_1,\cdots y_1^{(r_1-1)},\cdots,y_p,\dot{y}_p,\cdots,y_1^{(r_p-1)}]^T \quad (5)$$

The expression of the new system after coordinate transformation is

$$\dot{\zeta} = \hat{A}_1\zeta+\hat{A}_2\sigma+\hat{B}_1u$$

$$\dot{\sigma} = \hat{A}_3\zeta+\hat{A}_4\sigma+\hat{B}_2u$$

where $\hat{A}=\begin{bmatrix}\hat{A}_1 & \hat{A}_2\\ \hat{A}_3 & \hat{A}_4\end{bmatrix}=T^{-1}AT$ and $\hat{B}=\begin{bmatrix}\hat{B}_1\\ \hat{B}_2\end{bmatrix}$. Replacing $x(t)$ in (4) with the transformed dynamics, the control law to maintain the exact tracking can be written as

$$u_f=B_y^{-1}[y_d^{(r)}-A_\zeta\zeta(t)-A_\sigma\sigma(t)] \quad (6)$$

where

$$[A_\zeta \quad A_\sigma]=A_xT$$

internal dynamics can now be expressed as

$$\dot{\sigma} = \hat{A}_3\zeta+\hat{A}_4\sigma+\hat{B}_2B_y^{-1}[y_d^{(r)}-A_\zeta\zeta(t)-A_\sigma\sigma(t)]$$

$$= \hat{A}_\sigma\sigma(t)+\hat{B}_\sigma Y \quad (7)$$

where

$$\hat{A}_\sigma = \hat{A}_4-\hat{B}_2B_y^{-1}A_\sigma$$

$$\hat{B}_\sigma = [(\hat{A}_3-\hat{B}_2B_y^{-1}A_\zeta) \quad \hat{B}_2B_y^{-1}] \text{ and}$$

$$Y = [\zeta^T \quad y_d^{(r)T}]^T$$

in the same respect, equation (4) can now be written as

$$u(t) = B_y^{-1}[y_d^{(r)}-A_\zeta\zeta(t)-A_\sigma\sigma(t)]$$

$$= -B_y^{-1}A_\sigma\sigma(t)-[B_y^{-1}A_\zeta \quad -B_y^{-1}]Y$$

$$= \hat{C}_\sigma\sigma(t)+\hat{D}_YY(t) \quad (8)$$

where

$$\hat{C}_\sigma = -B_y^{-1}A_\sigma \text{ and}$$

$$\hat{D}_Y = -[B_y^{-1}A_\zeta \quad -B_y^{-1}]$$

Equation (7) together with equation (8) form the inverse system and can be represented in state space form

$$\dot{\sigma}(t) = \hat{A}_\sigma\sigma(t)+\hat{B}_\sigma Y(t) \quad (9)$$

$$u(t) = \hat{C}_\sigma\sigma(t)+\hat{D}_YY(t) \quad (10)$$

and represented as in Figure 2(see Appendix A.4 for details of matrices A_σ,B_σ,C_σ,D_Y).

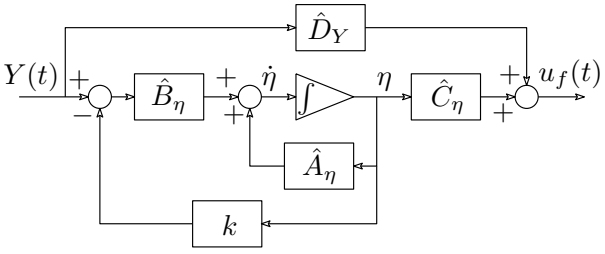


Figure 2: Block diagram of the inverse system

4 Inverting the manipulator

The linear model has 17 states distributed as:

- $x_1(t)=i_1(t)$
- $x_2(t)=w_{11}(t)$
- $x_3(t)=\dot{w}_{11}(t)$
- $x_4(t)=w_{12}(t)$
- $x_5(t)=\dot{w}_{12}(t)$
- $x_6(t)=w_{21}(t)$
- $x_7(t)=\dot{w}_{21}(t)$
- $x_8(t)=w_{22}(t)$
- $x_9(t)=\dot{w}_{22}(t)$
- $x_{10}(t)=\theta_1(t)$
- $x_{11}(t)=\dot{\theta}_1(t)$
- $x_{12}(t)=\theta_2(t)$
- $x_{13}(t)=\dot{\theta}_2(t)$
- $x_{14}(t)=\theta_3(t)$
- $x_{15}(t)=\dot{\theta}_3(t)$
- $x_{16}(t)=i_3(t)$
- $x_{17}(t)=i_2(t)$

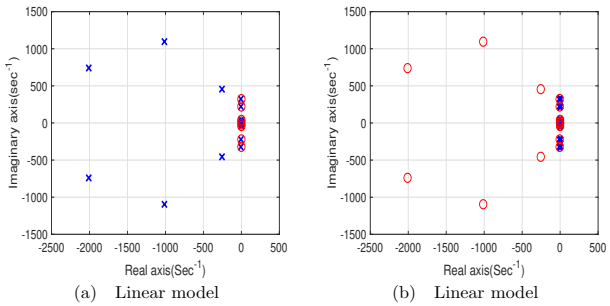
where i_j denotes the armature current to the servomotor driving joint $j(j=1,2,3)$, θ_j and $\dot{\theta}_j$ are the instantaneous joint angles and joint velocities of joint $(j=1,2,3)$, respectively, whereas (w_{11},w_{12}), (w_{21},w_{22}) and their derivatives denote the flexure variable for links 1 and 2 respectively.

Remark 1. *In the modelling of the manipulator in Maplesim, the lengths of links 1 and 2 are broken into two to accommodate an instrument to measure the strain. In regard to this, in the linearized model, the flexure variable has two parts as w_{11},w_{12} for link 1 and w_{21},w_{22} for link 2. Except for having twice as many flexure variables as the number of links, breaking the links does not affect the performance of the model.*

With a relative degree of $r=(3,3,3)$, the internal dynamics, $\sigma(t)$, were taken as the flexure variables, whereas the output variables and their derivatives ζ, were taken as the three joint angles, velocities and motor currents, *i.e.*

$$\sigma(t)=\begin{bmatrix}x_2(t)\\x_3(t)\\x_4(t)\\x_5(t)\\x_6(t)\\x_7(t)\\x_8(t)\\x_9(t)\end{bmatrix}=\begin{bmatrix}w_{11}(t)\\\dot{w}_{11}(t)\\w_{12}(t)\\\dot{w}_{12}(t)\\w_{21}(t)\\\dot{w}_{21}(t)\\w_{22}(t)\\\dot{w}_{22}(t)\end{bmatrix},\quad \zeta(t)=\begin{bmatrix}x_1(t)\\x_{10}(t)\\x_{11}(t)\\x_{12}(t)\\x_{13}(t)\\x_{14}(t)\\x_{15}(t)\\x_{16}(t)\\x_{17}(t)\end{bmatrix}=\begin{bmatrix}\theta_1(t)\\\dot{\theta}_1(t)\\\ddot{\theta}_1(t)\\\theta_2(t)\\\dot{\theta}_2(t)\\\ddot{\theta}_2(t)\\\theta_3(t)\\\dot{\theta}_3(t)\\\ddot{\theta}_3(t)\end{bmatrix}$$

Poles and zeros of the linear model and its inverse are as shown below



(a) Linear model (b) Linear model

Figure 3: Poles and zeros of the plant and its inverse

Upon inversion, the eight poles corresponded to the internal dynamics, i.e. the flexure variables. These poles were found to be lying on the imaginary axis implying that these variables are marginally stable and that they will not decay to zero with time. This was addressed by applying pole placement technique to slightly shift these poles to the right. This is important as it means that the stability of the internal dynamics and the resulting inverse is assured. Also the stable model still remains to be the inverse of the linear model. Consequently the stable solution of the internal dynamics follows from solving the ODE in equation 9 as

$$\sigma(t) = e^{A_\sigma t}\sigma(0)+\int_0^t e^{A_\sigma(t-\tau)}B_\sigma Y(\tau)d\tau \quad (11)$$

where τ is a dummy variable. The first term represent the zero-input response whereas the second term is the zero-state response. With negative eigenvalues of the matrix A_σ, which is enforced using pole placement technique, it can be deduced from this expression that the internal dynamics $\sigma(t)$ are bounded for bounded external dynamics $Y(t)$.

Driving the manipulator via an inverse controller implies that the joint angles will follow the desired trajectories exactly. For the high speed operations involving step or square wave trajectories; however, joint velocities during the rising and the falling edges would be too high thus not safe for the operators. It can also lead to mechanical failures. This was addressed by introducing second order bilinear low pass filter, before the inverse controller, of the form

$$f(s)=\frac{1}{(\lambda s+1)^2}$$

where λ is an adjustable parameter for limiting the manipulator speeds to safe levels.

Remark 2. *Operation without filter leads to very high speeds, hence exposing the operator and environment to risk, also risking the mechanical well being of the manipulator.*

Remark 3. *Operation with filters without the inverse controller means that the high frequency components of the trajectories are removed leading to a very high joint error.*

Remark 4. *The inverse controller ensures that the joint trajectories follows the desired trajectory, the filter ensure safe operation speeds.*

5 Direct strain feedback control

The theory of Direct Strain Feedback (DSFB) was developed by Luo [22, 23] and experimented with a one-link flexible manipulator. In this control scheme, the strain measured at the root of the flexible link is multiplied by a constant gain k and the resultant signal is used to modify the control law as a negative feedback. The overall effect of direct strain feedback is to increase the system damping coefficient thereby leading to a rapid decay of transverse vibrations and torsional vibration as a result of coupling between the two types of vibrations. In [22], the author shows that the technique can satisfactorily dampen link vibrations. He also analytically derived the proof that the resulting closed loop is asymptotically stable. A block diagram showing the hybrid of the filtered inverse feedforward controller and the DSFB is shown in Figure 4.

From the figure, the new control law is expressed as

$$u(t)=\theta_f(t)-\theta(t)-k\varepsilon(0,t)$$

where:

$\theta_f(t)$ - Trajectory tracking signal generated by the filtered inverse controller
$\theta(t)$ - Joint angle of the flexible manipulator
$\varepsilon(0,t)$ - Strain at the root of the links
k - Strain feedback controller gain

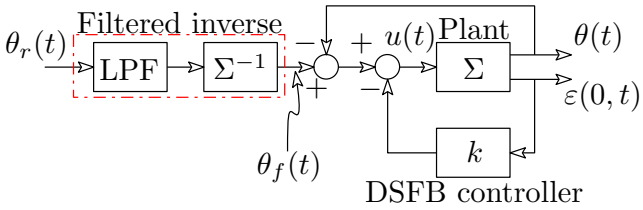


Figure 4: Hybrid control setup

6 Results and Discussion

6.1 Modelling and validation results

To validate the model developed and linearized in section 2, simulations were conducted in Matlab Simulink on the linearized models and the performance of the model compared with the existing flexible manipulator. The task involved moving the joints as in a simple pick and place task popular in industries e.g soldering, painting etc.

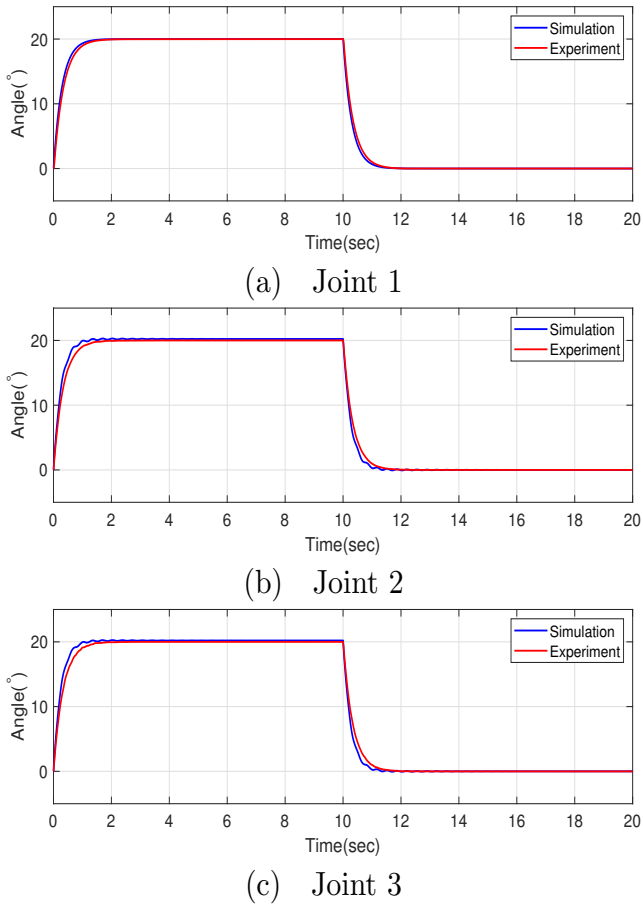


Figure 5: Comparison of Joint angles between linear and the actual manipulator

Figure 5 shows the joint trajectories for joint 1, 2 and 3 for the model and the manipulator. From the figures, we can see a perfect agreement of the model with the existing manipulator.

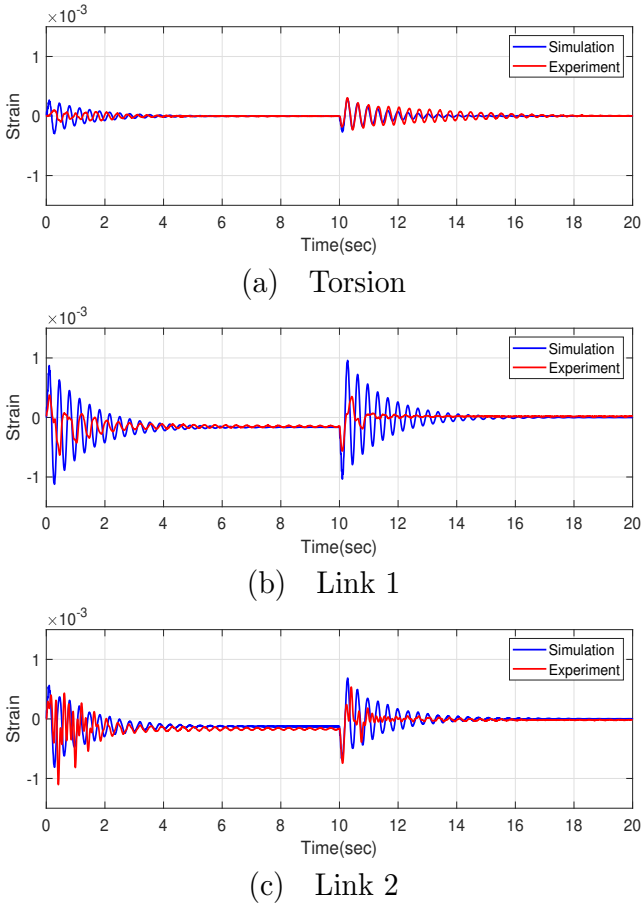


Figure 6: Comparison of strain between linear and the actual manipulator

Figure 6 validates the model employed in this work in terms of the vibrations excited. It shows the strain information of the linear model against the actual manipulator. Perfect agreement between joint angles in torsional and links strain in 6(a-c) can be observed. Observations in figures 5 and 6 leads to the deduction that the linear model represents an accurate model of the manipulator. In addition, the inverse derived from the linearized model represent an accurate inverse of the actual manipulator.

6.2 Simulation results

This section presents simulation results of the validated model subjected to the inverse controller, the strain feedback controller and a hybrid controller of the two. Simulations were conducted in Matlab/Simulink for a desired joint trajectory that involved moving the joints at an angle of 20 degrees for 10 seconds and back to the vertical position for 10 more seconds. Typical application s of such trajectories are in soldering and other pick and place related tasks. Simulink model of the simulation setup is as shown in Figure 7.

The servomotors used here were of the speed reference type. Thus, the angle feedback formed the outer feedback loop, having unity feedback gain, while strain feedback was in the inner loop with a feedback gain $k = 0.4$. To avoid errors due to self-weight, the gravitational effect is compensated for in the strain

signal. The compensation was done by removing the self-weight offset before being fed back through the DSFB controller.

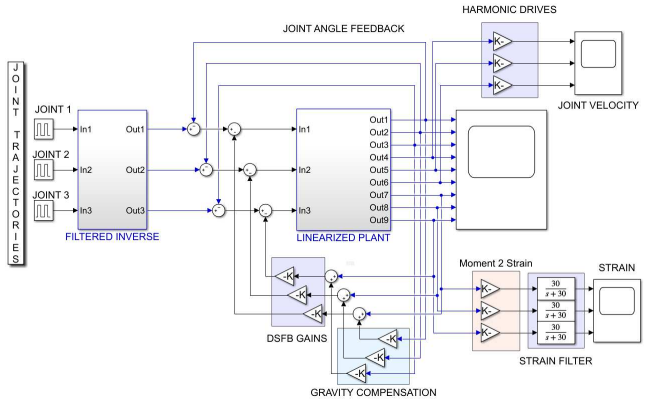
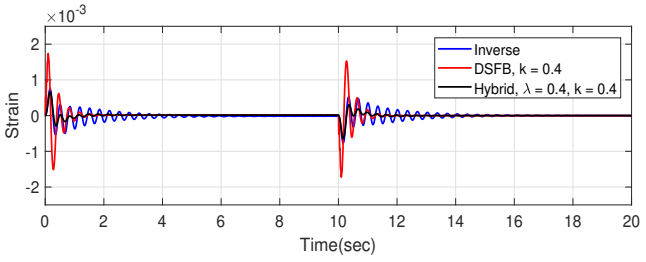
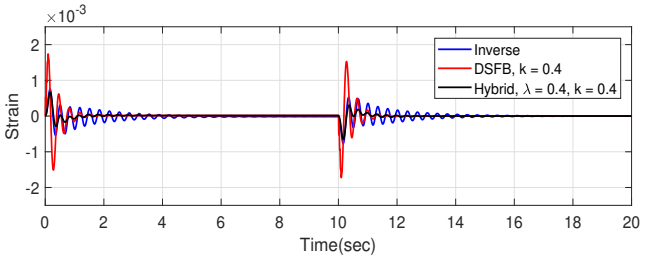


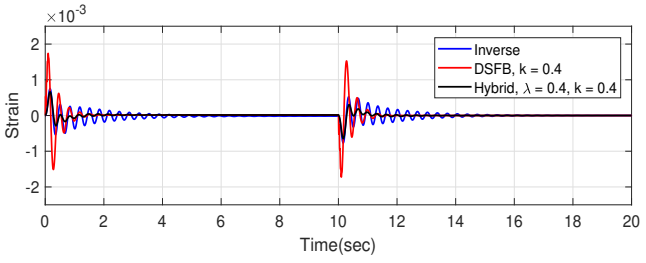
Figure 7: Simulation setup in MATLAB



(a) Link 1, torsion



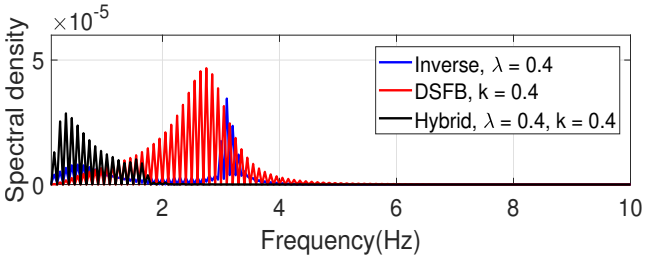
(b) Link 1, in plane



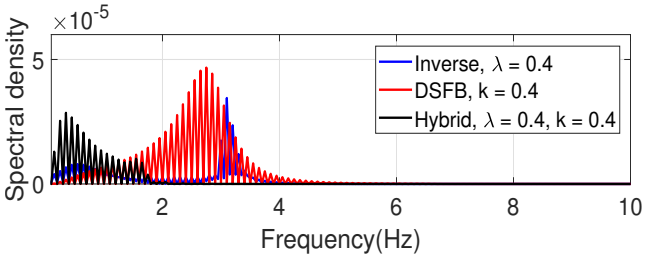
(c) Link 1, in plane

Figure 8: Strain information of the linear model

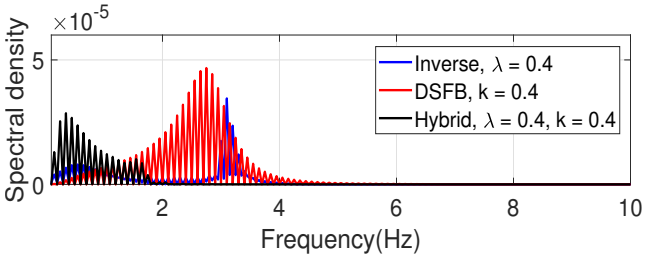
From the figures of torsional strain(Figure 8a), in plane strain for link 1(Figure 8b) and link 2(8c), it can be seen that the inverse controller had an upper hand in suppressing transient vibrations caused by the sudden starting and sudden stopping. However, these vibrations lasted for a relatively longer time. Interestingly, DSFB, though it had very poor transient response, it was very strong in dealing with residue vibrations. The hybrid of the two controller was inherently better in handling both the transient and the residue vibrations, hence, outperformed the individual controllers. Figure 9 gives a pictorial evidence of the of the comparison and the strengths of the individual controllers.



(a) Link 1, torsion



(b) Link 1, in plane



(c) Link 1, in plane

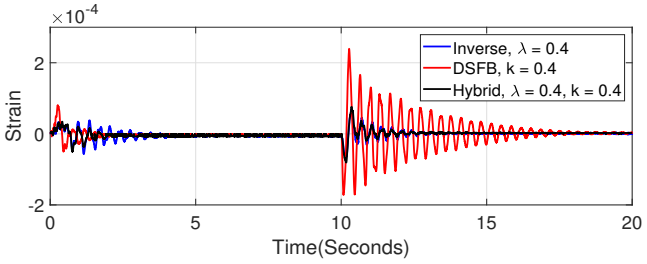
Figure 9: Strain power spectrum density

6.3 Experimental results

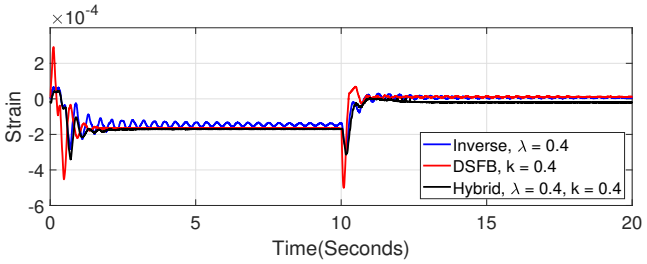
The experiment involved moving the three joints at an angle of 20 degrees, using a step signal lasting for 10 seconds, followed by returning to its original vertical position for another 10 seconds for a case without tip mass and with a tip mass of 100g. Strain measurement was achieved by attaching strain gauges at the root of respective links for torsional, link 1 in plane and link 2 in plane strain. Figure 10 shows the experimental setup of this work.

Figure 10: Experiment setup

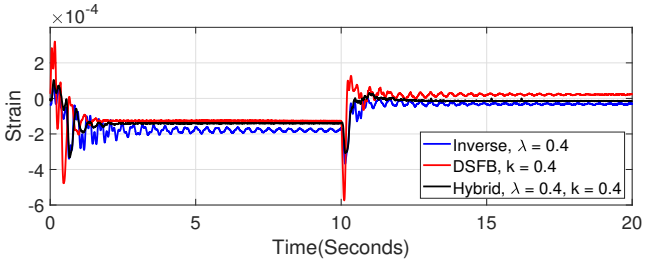
Figure 11 shows the experimental results for the torsional strain, 11a, and strain for the two links, (11b, 11c) for a manipulator without any tip load. The figures show the comparison of the inverse controller, the DSFB controller and a hybrid of the two. It can be seen that the inverse controller had an upper hand in dealing with motion induced vibrations. On the other hand, the DSFB ensured a rapid decay of the residue vibrations. A combination of the two controllers as a two degree of freedom controller, yielded a system characterised by minimal motion-induced vibrations decaying very rapidly. The hybrid of the two yielded minimal strain amplitudes and shortened the duration of the vibrations. Actually, the overall performance of the hybrid was better than that of the controllers in their areas of strengths, individually. This is attributed to the combined strengths of the two controllers.



(a) Link 1, torsion



(b) Link 1, in plane

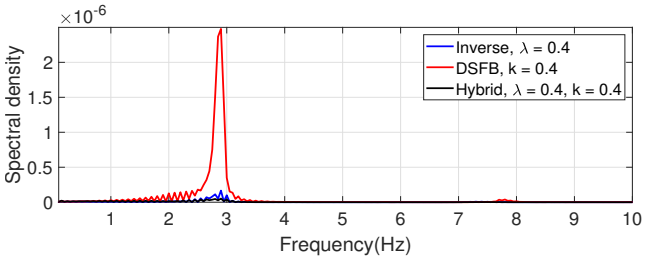


(c) Link 1, in plane

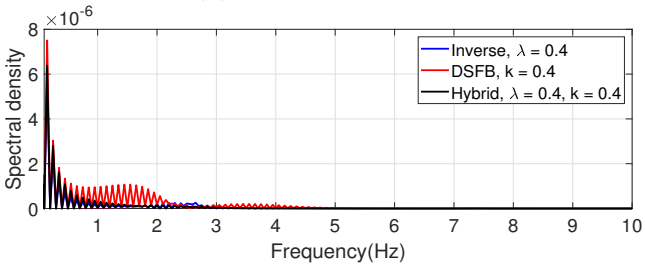
Figure 11: Torsional and lateral strain without any load

In plane strain in the first 10 seconds of the links 1 and 2(11b, 11c), an offset from zero strain can be seen. This error is associated with the fact that, during this time, links 1 and 2 are tilted and thus affected by gravity. Consequently, the bending strain at the root of the links did not converge to zero but remained to be a value of the distortion due to the self weight of the links for the entire period. However, this effect doesn't affect torsional strain.

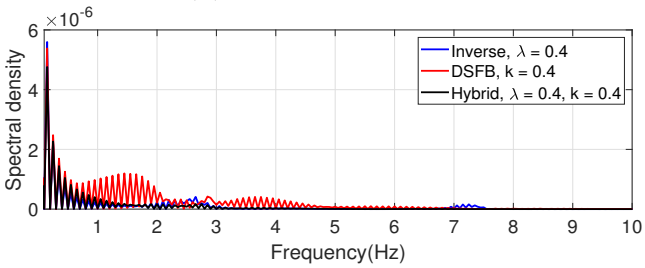
Figure 12 shows the strain spectral power density without any load, where the improvement was very significant. It can be seen that individual controllers suppress different frequencies while the hybrid inherits these capabilities, better yet outperforming the individual controllers. Again, the effect of self weight due to the tilted status of links 1 and 2 can be seen in the lower part of the spectrum (Figures 12b,12c), but it is absent in the spectrum for torsional strain (12a).



(a) Link 1, torsion



(b) Link 1, in plane

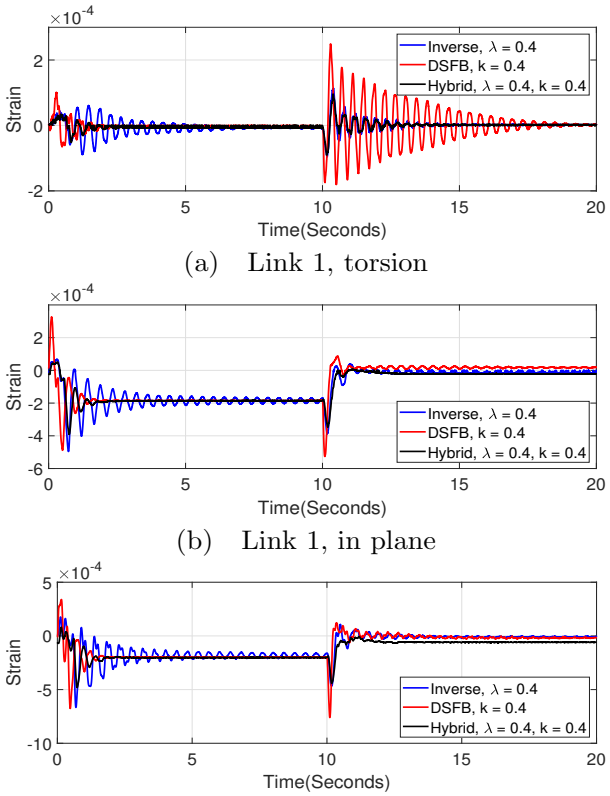


(c) Link 1, in plane

Figure 12: Strain spectral density without any load

With a load of 100g attached at the distal end of link 2, Figure 13 shows the torsional strain(13a) and in plane lateral strain for links 1 and 2(13b,13c). From the figure, the vibrations are a bit severe relative to those experience in a system without load. This is attributed to the fact that loading a flexible manipulator leads to excitation of more severe vibration at a relatively lower frequency when compared to a manipulator without any load. The offset due to self weight imposed by gravity during the first 10 seconds is also relatively higher.

The performance of the inverse controller is commendable in attenuating link vibration, however, complete mitigation is not possible and severe residues remains for the entire duration of operation. Strain feedback on the other hand, though its performance in eradicating the residues is very good, it is poor in dealing with trajectory induced vibrations. The hybrid of the two controllers, having inherited the complementary strengths of each controller is able to deal with both transient and residue vibrations.

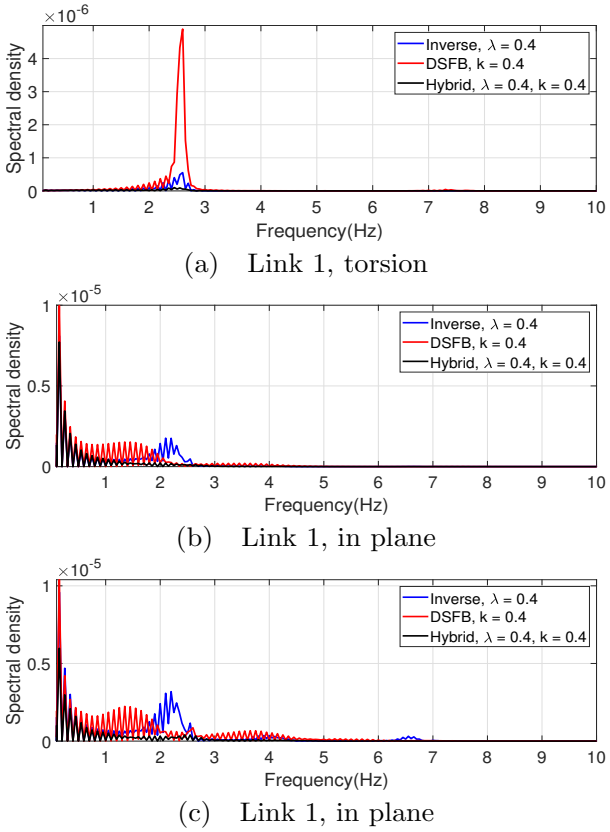


(a) Link 1, torsion

(b) Link 1, in plane

(c) Link 1, in plane

Figure 13: Torsional and lateral strain with a load of 100g



(a) Link 1, torsion

(b) Link 1, in plane

(c) Link 1, in plane

Figure 14: Strain spectral density with a load of 100g

To investigate the effect of loading on the strain spectral power density, Figure 14 illustrates the spectrum for torsional strain(14a) and in plane lateral strain(14b, 14c) for links 1 and 2 respectively. Comparing these frequency responses, we can observe that: 1). The peaks are higher than those seen in a system without any load, and 2). these peaks occurred at a slightly lower frequencies.

7 Conclusion

We successfully developed, linearized and validated a model of a 3D, two link, flexible manipulator. A stable right inverse of the linear model was developed, augmented with a low pass filter, and used as a feedforward controller. Together with a direct strain feedback controller, $k = 0.4$, in the feedback loop, the combination formed a two degree of freedom controller. From both simulations and experimental results, we found the inverse controller has an upper hand in handling motion induced vibrations which arose from sudden starting and stopping. Also, the strain feedback controller ensures rapid decay of the residue vibrations. Finally, the hybrid of the two controllers inherently has the strengths of the two controllers, exhibited a superior performance of suppressing both the transient and the residue vibrations.

References

[1] Waweru Njeri, Minoru Sasaki, and Kojiro Matsushita, "Two-degree-of-freedom control of a multilink flexible manipulator using filtered inverse feedforward controller and strain feedback controller," in *2018 IEEE International Conference on Applied System Invention (ICASI)*, pp. 972–975, April 2018.

[2] Y. Bian, Z. Gao, X. Lv, and M. Fan, "Theoretical and experimental study on vibration control of flexible manipulator based on internal resonance," *Journal of Vibration and Control*, vol. 0, no. 0, p. 1077546317704792, 2017. doi:10.1177/1077546317704792.

[3] Y. Bian and Z. Gao, "Nonlinear vibration control for flexible manipulator using 1: 1 internal resonance absorber," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 0, no. 0, p. 1461348418765951, 2018. doi:10.1177/1461348418765951.

[4] M. H. Shaheed and O. Tokhi, "Adaptive closed-loop control of a single-link flexible manipulator," *Journal of Vibration and Control*, vol. 19, no. 13, pp. 2068–2080, 2013. doi:10.1177/1077546312453066.

[5] Y. Wang, Q. Zheng, H. Zhang, and L. Miao, "Adaptive control and predictive control for torsional vibration suppression in helicopter/engine system," *IEEE Access*, vol. 6, pp. 23896–23906, 2018.

[6] B. Luo, H. Huang, J. Shan, and H. Nishimura, "Active vibration control of flexible manipulator using auto disturbance rejection and input shaping," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 228, no. 10, pp. 1909–1922, 2014. doi:10.1177/0954410013505951.

[7] Z. Masoud, M. Nazzal, and K. Alhazza, "Multimode input shaping control of flexible robotic manipulators using frequency-modulation.," *Jordan Journal of Mechanical & Industrial Engineering*, vol. 10, no. 3, 2016.

[8] Z. Mohamed and M. O. Tokhi, "Vibration control of a single-link flexible manipulator using command shaping techniques," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 216, no. 2, pp. 191–210, 2002. doi:10.1243/0959651021541552.

[9] L. Malgaca, ahin Yavuz, M. Akda, and H. Karaglle, "Residual vibration control of a single-link flexible curved manipulator," *Simulation Modelling Practice and Theory*, vol. 67, pp. 155 – 170, 2016.

[10] . Yavuz, L. Malgaca, and H. Karaglle, "Vibration control of a single-link flexible composite manipulator," *Composite Structures*, vol. 140, pp. 684 – 691, 2016.

[11] X. Zhang, J. K. Mills, and W. L. Cleghorn, "Experimental implementation on vibration mode control of a moving 3-prr flexible parallel manipulator with multiple pzt transducers," *Journal of Vibration and Control*, vol. 16, no. 13, pp. 2035–2054, 2010. doi:10.1177/1077546309339439.

[12] S. O. Reza Moheimani, "A survey of recent innovations in vibration damping and control using shunted piezoelectric transducers," *IEEE Trans. Control Syst. Technol*, pp. 482–494, 2003.

[13] G. Song, S. P. Schmidt, and B. N. Agrawal, "Active vibration suppression of a flexible structure using smart material and a modular control patch," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 214, no. 4, pp. 217–229, 2000. doi:10.1243/0954410001532024.

[14] Agrawal B. N. and Bang H., "Active vibration control of flexible space structures by using piezoelectric sensors and actuators," *In Proceedings of 14th Biennial ASME Conference*, Sept. 1993.

[15] M. L. Kerr, S. Jayasuriya, and S. F. Asokanthan, "Qft based robust control of a single-link flexible manipulator," *Journal of Vibration and Control*, vol. 13, no. 1, pp. 3–27, 2007. doi:10.1177/1077546306064826.

[16] D. Halim and S. O. R. Moheimani, "Spatial resonant control of flexible structures-application to a piezoelectric laminate beam," *IEEE Transactions on Control Systems Technology*, vol. 9, pp. 37–53, Jan 2001.

[17] M. Sasaki, A. Asai, T. Shimizu, and S. Ito, "Self-tuning control of a two-link flexible manipulator using neural networks," in *2009 ICCAS-SICE*, pp. 2468–2473, Aug 2009.

[18] Waweru Njeri, Minoru Sasaki, and Kojiro Matsushita, "Enhanced vibration control of a multilink flexible manipulator using filtered inverse controller," *ROBOMECH Journal*, vol. 5, p. 28, Nov. 2018.

[19] J. Zhang, L. He, E. Wang, and R. Gao, "Active vibration control of flexible structures using piezoelectric materials," in *2009 International Conference on Advanced Computer Control*, pp. 540–545, Jan 2009.

[20] F. Cao and J. Liu, "Vibration control for a rigid-flexible manipulator with full state constraints via barrier lyapunov function," *Journal of Sound and Vibration*, vol. 406, pp. 237 – 252, 2017.

[21] D. Liberzon, A. S. Morse, and E. D. Sontag, "Output-input stability and minimum-phase nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 47, pp. 422–436, Mar 2002.

[22] Z.-H. Luo, "Direct strain feedback control of flexible robot arms: new theoretical and experimental results," *IEEE Transactions on Automatic Control*, vol. 38, pp. 1610–1622, Nov 1993.

[23] Z.-H. Luo and Y. Sakawa, "Gain adaptive direct strain feedback control of flexible robot arms," in *TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on*, vol. 4, pp. 199–202 vol.4, Oct 1993.

A Appendix

A.1 Manipulator specifications

Table 1: Specifications of the flexible manipulator

Servo moter1	(Joint1)	Type	V850-012EL8
	Rated armature voltage	80	V
	Rated armature current	7.6	A
	Rated power	500	W
	Rated spindle speed	2500	rpm
	Rated torque	1.96	N.m
	Moment of inertia	6×10^{-4}	kg.m^2
	Mass	4.0	Kg
Servo moter2	(Joint2)	Type	T511-012EL8
	Rated armature voltage	75	V
	Rated armature current	2	A
	Rated power	100	W
	Rated spindle speed	3000	rpm
	Rated torque	0.34	N.m
	Moment of inertia	3.7×10^{-5}	kg.m^2
	Mass	0.95	Kg
Servo moter3	(Joint3)	Type	V404-012EL8
	Rated armature voltage	72	V
	Rated armature current	1	A
	Rated power	40	W
	Rated spindle speed	3000	rpm
	Rated torque	0.13	N.m
	Moment of inertia	8.4×10^{-6}	kg.m^2
	Mass	0.4	Kg
Encoder	Reduction ratio	1/100	P/R
	Spring constant	1.6×10^{4}	Nm/rad
Harmonic drive -joint1	Type	CSF-40-100-2A-R-SP	
	Reduction ratio	1/100	
	Spring constant	23	Nm/rad
	Moment of inertia	4.50×10^{-4}	kg.m^2
Harmonic drive -joint2	Type	CSF-17-100-2A-R-SP	
	Reduction ratio	1/100	
	Spring constant	1.6×10^{-4}	Nm/rad
	Moment of inertia	7.9×10^{-6}	kg.m^2
Harmonic drive -joint3	Type	CSF-14-100-2A-R-SP	
	Reduction ratio	1/100	
	Spring constant	7.1×10^{-5}	Nm/rad
	Moment of inertia	3.3×10^{-6}	kg.m^2
Link1	Material	Stainless steel	
	Length	0.44	m
	Radius	5×10^{-3}	m
Link2	Material	Aluminum	
	Length	0.44	m
	Radius	4×10^{-3}	m
Strain Gauge	Type	KGF-2-120-C1-23L1M2R	

A.2 Maplesim model

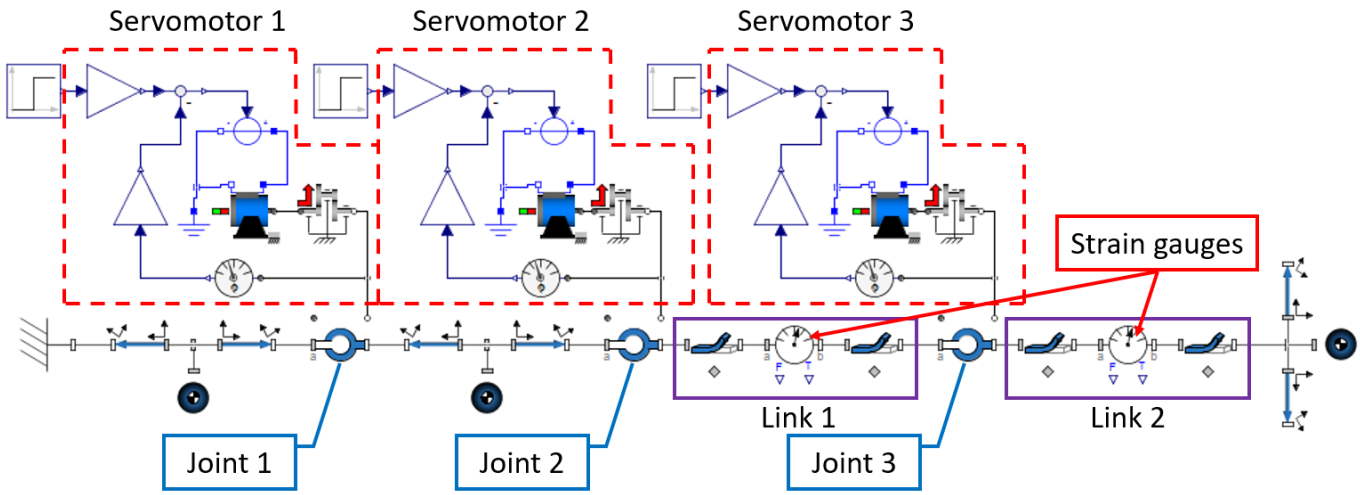


Figure 15: Maplesim model of the manipulator

Table 2: Building block of the arm in Maplesim

S.No.	Symbol	Component name	Short description.
1.	pin_an pin_ap flange	pmdc motor	Models a DC Machine with permanent magnets.
2.	y	Step	Generates a real step signal with variable height.
3.	flange_a flange_b frame_a frame_b	Revolute	Joint allowing one rotational degree of freedom about a given axis.
4.	flange_a RO	Angle Sensor	Measures the absolute flange angle.
5.	frame_a frame_b FMOutputF FMOutputT	Force and Moment	Measures and outputs the forces and moments acting between two frames.
6.	frame_a	Rigid bode	Center of mass frame with associated mass and inertia matrix.
7.	frame_a frame_b base	Flexible Beam	A flexible beam with axial, lateral, and torsional deformations.
8.	u y	Gain	Outputs the product of a gain value with the input signal.
9.	flange_a flange_b	Lossy Gear	Gearbox with mesh efficiency and bearing friction. Represent the harmonic drive.
10.	frame_a frame_b	Rigid Body Frame	Frame with a fixed displacement and orientation relative to a rigid body center of mass frame.

A.3 State, input, output and the transfer matrices of the linearized model

$$A = \begin{bmatrix}
-509.09 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 46143.66 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -8955.97 & 0 & -503.85 & 0 & 7341.21 & 0 & 963.6405 & 0 & 0 & 0 & -58.54 & -13.29 & 12.57 & -0.22 & -0.55 & -38.84 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -23638.29 & 0 & -105473.64 & 0 & 36272.29 & 0 & 107151.69 & 0 & 0 & 0 & -8.25 & -1.87 & 802.64 & 2948.43 & 7350.05 & -5.46 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 57299.95 & 0 & 5944.79 & 0 & -60522.46 & 0 & -7988.99 & 0 & 0 & 0 & 453.84 & 508.15 & -100.48 & 3.1590 & 7.87 & 1484.68 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3910.71 & 0 & 9243.10 & 0 & -4195.63 & 0 & -12690.78 & 0 & 0 & 0 & 0.90 & 0.20 & -69.64 & -35.6798 & -88.94 & 0.60 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-5.89 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1.74 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & -237.96 & 0 & -0.70 & 0 & 1161.67 & 0 & 0.89 & 0 & 0 & 0 & 0.12 & -40.47 & 0.01 & -0.0004 & -0.0009 & -118.26 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & -3.89 & 0 & 1106.14 & 0 & 7.22 & 0 & -155.35 & 0 & 0 & 0 & -0.0017 & -0.0004 & 0.17 & -177.77 & -443.16 & -0.0011 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8903.10 & -3857.14 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 18260.93 & 0 & 0 & 0 & -2000
\end{bmatrix}$$

$$B = \begin{bmatrix}
-909.0909091 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & -238.0952381 \\
0 & -416.6666667 & 0
\end{bmatrix}$$

$$C = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 57.2958 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 57.2958 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 57.2958 & 0 & 0 & 0
\end{bmatrix}$$

$$D = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}$$

A.4 State, input, output and the transfer matrices of the internal dynamics of the inverse model

$$\hat{A}_\sigma = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -8877.8083 & 0 & -505.0084 & 0 & 6959.6426 & 0 & 963.5402 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -23691.8248 & 0 & -87127.7174 & 0 & 36338.2298 & 0 & 104574.9609 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 54312.6517 & 0 & 5955.6170 & 0 & -45939.3525 & 0 & -7980.4902 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 3910.2812 & 0 & 9021.0946 & 0 & -4191.1701 & 0 & -12659.5994 & 0 \end{bmatrix}$$

$$\hat{B}_\sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1.0225 & 0 & 0.2194 & 0 & 0 & 0 & 0.0057 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.1445 & 0 & 14.0596 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7.9491 & 0 & -1.7512 & 0 & 0 & 0 & -0.21900 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0158 & 0 & -1.2161 & 0 & 0 & 0 & 0 & 0 & 0.0035 & 0 & 0 & 0 \end{bmatrix}$$

$$\hat{C}_\sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -9.6577 & -0.0048 & -0.0285 & 0 & 47.1468 & 0.0235 & 0.0364 & 0 \\ -0.1422 & 0 & 40.4355 & 0.0104 & 0.2636 & 0 & -5.6792 & -0.0014 \end{bmatrix}$$

$$\hat{D}_\sigma = \begin{bmatrix} 0 & -1 & 0 & 0 & -0.8887 & 0 & 0 & -0.0016 & 0 & 0 & 0 & 0 \\ -0.9999 & 0 & 0 & -0.7935 & 0 & 0 & -0.0007 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.9998 & 0 & 0 & -0.7660 & 0 & 0 & -0.0006 & 0 & 0 & 0 \end{bmatrix}$$

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 270-279 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Enabling the Edge - A method for dynamic virtualizable connections for 5G deployments

James Robert Pogge[*,1], Stephen lucien Scott[2]

[1]*Electronics Computer Engineering Dept, Tennessee Technological University, 38505, USA*

[2]*Computer Science Dept, Tennessee Technological University, 38505, USA*

ARTICLEINFO

Article history:
Received: 20 December, 2018
Accepted: 25 March, 2019
Online: 10 April, 2019

Keywords:
5G
Terabit
Edge routing

ABSTRACT

The telecommunications industry continues to move forward with plans for 5G service with rates including 6Gb/s followed by upgrade paths promising 20Gb/s. Despite advances in wireless technology associated with 5G roll-outs, there remains the problem of providing sufficient back-haul throughput and efficient handling of hand-off services within the relatively small projected cell boundaries that can handle the higher frequencies required for these data rates. Dense network discussions include proposals for macrocell centric groupings that directly support multiple micro-cells to create a larger service area. The two-tiered approach presents additional complexities when considering the dynamic nature of mobile wireless networks. The added delays of multi-tier hand-offs compound the already heavily burdened throughput capability of the back-haul service towers. The 5G enabled network should easily allow legacy technology 4G and 3G connectivity without sacrificing usable bandwidth by dedicating limited back-haul ports to legacy technology. The presented methodology is a hybrid distributed optical switch embedded directly within the micro-cell towers that simplify hand-off and allows for dynamic allocation of bandwidth on demand. Simulation results and a test-bed optical network interface suggests that bandwidth on demand topology is possible combined with software-defined dynamic routing in a flat plane topology. The solution uses already available technology found in high-speed Internet backbone distributions.

1 Introduction

Large scale deployments of 5th generation (5G) enabled towers to require ever-increasing access to bandwidth to provide near real-time service to connected customers, [1]. At the high frequencies projected for 5G deployments 56-60GHz, the effects of propagation fade and multi-path issues in dense urban environments become more prevalent. [2]. [3]. The pre-rollout solutions being tested now trend toward a network model of multisector closely spaced cell towers, [4] , [5]. The emerging deployment model is classified as an ultra-dense network, [6], one that is tightly spaced clusters of overlapped minicells. The emerging 5G topologies include multiple backhaul proposals using millimeter wave macro cells as shown in figure 1, a centralized aggregated service portal with fiber access switch to the Internet [7]. These macrocells use millimeter-wave wireless modems within the higher capacity 28-100 GHz proposed RF bands. The bandwidth requirements of the microcells stress the limit even for 60 GHz channels. High-density RF modulation directly from the macro tower to the microcell service cluster using OFDM (Orthogonal Frequency Division Multiplexing) is necessary to deliver promised service to multiple wireless customers while maintaining regulatory requirements for supporting legacy 3G and 4G LTE access, [7].

In existing cell networks such as those found in 4G Long Term Evolution(LTE) base stations, the moderate backhaul requirements are provided through the proper arrangement of high-speed network switches, [8], typically a series of aggregated gigabit Ethernet. In high-density networks of this type, the switch latency

[*]Corresponding Author: James Robert Pogge, Kingstom, TN USA, (865) 335-4164 & jrpogge42@students.tntech.edu

https://dx.doi.org/10.25046/aj040235

and density of data are limited by existing centralized 10G switches. The high volume of traffic times makes 5G promised real-time data streams difficult to achieve. Consider that the proposed 5G service rates of nominally 5Gb/s to the projected 20 Gb/s service require high bandwidth when each sector microcell should service 10 or more customers to meet market expectations.

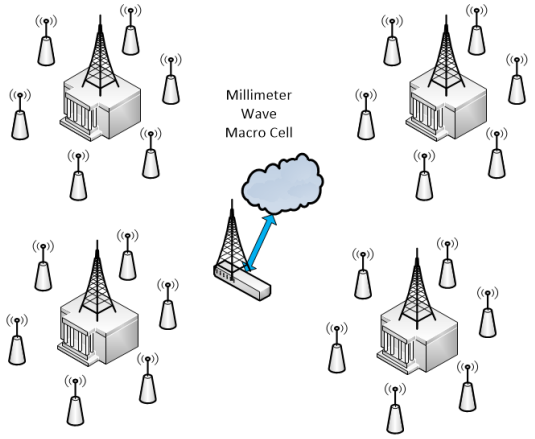


Figure 1: Millimeter wave Macrocell Backhaul

In a dense urban environment that demand load can change dynamically, [8]. In the latest industry projection for data usage requirements, the demographic shift of post-millennials shows that tablets are gaining in favor. Video phone calls are displacing texting and traditional voice calls especially in urban work environments. The current trends suggest that the majority connection will use a significant portion of the advertised 6Gb/s connections, [8]. Despite Multiple Input, Multiple Output (MIMO) radios covering the last connection in the microcell, the service to individual users is determined by the underlying switch topology under stress in high-density population centers and not the carrying capacity of the local radio modem, [9]. Additionally, the cost of a multi-tiered TCP/IP switch network in smaller area ultra-dense networks due to the propagation limits of the proposed RF bands implies that following the existing 4G aggregated network model becomes prohibitively expensive. [10], [11].

In this paper, the authors propose a dynamically reconfigurable network interface that allows bandwidth allocation to be resolved on the fly and shared equally with adjacently connected towers. The backhaul demand is accomplished through a virtualizable plane parallel optical topology that relies on the use of available fiber optic technology originally intended for long distance backhaul Internet applications. The proposed solution includes emerging flex-net transceivers located directly in the minicell sector 5G radios eliminating the need for expensive aggregated switches. The individual sectored wireless links gain direct access to the entire optical bandwidth of the backhaul network as required without upgrading micro cell towers. All of the microcell sector RF modems are serviced directly through a locally centrally managed sector scheduler node. This proposal includes a method for (SDN) software-defined network bandwidth allocation, use scheduling, and access control. Service scalability is handled through user-specified virtual networks connections that coexist with existing allocated fiber traffic. New connection and microcell commissioning can be achieved entirely without interference to existing network traffic flow or pathway allocations.

This paper introduces the concept of virtualizable ports or service interconnects by utilizing existing technology previously reserved for the highest speed terabit optical Internet backbone. The goal of this work is a presentation of a truly programmable SDN 5G solution. The result is a near flat network topology capable of accessing maximum available bandwidth on demand while maintaining system scalability. The embedded optical switch in the sector modem edge device is direct to fiber routing switch eliminating the need for a more traditional centralized layered network switch distribution approach.

2 Legacy Network Model

Typically in a switch based topology, increased access to bandwidth is limited by the availability of existing ports or pathways through the aggregated layered switches. In abnormal traffic situations breakdown occurs when the number of allocated switch ports exceed the expected number of users in a given sector. In the use case of a special event, such as a conference or sporting event where unusual traffic is concentrated in a smaller area, this requires the installation of additional switch devices installed specifically for the event along with increased numbers of temporary microcell sites. Additionally, connection delays in adjacent macrocell sectors are introduced as routing paths are arranged to handle the increased load. In ultra-dense networks, as expected the probability of achieving a successful connection to the user network decreases as the number of active nodes increase. Mitigating this effect requires dynamic allocation of bandwidth to the microcell. Responding to the increased demand during peak usage times is altered when unusual circumstances occur within the service area demographics.

The switch fabric connecting the tower to adjacent towers and subsequently the backhaul Internet connection must exceed the expected number of active end-user nodes without interfering with existing service connections. The probability that more than one node (N+1) attempts to access the network port simultaneously is expressed in Equation 1the probability (p) of access success is directly affected by the number of users attempting to connect. Which naturally implies that increased activity will adversely affect available bandwidth as the number of endpoint nodes (users) increases. The backhaul bandwidth is further reduced as the shared bandwidth of each microcell connection is assigned. Additionally, adjacent sectors need reliable

access to backhaul bandwidth to help take up the slack of the system and ensure timely hand-offs.

$$\sum_{i=1}^{N+1} \binom{N}{i} p^i (1-p)^{N-i} \tag{1}$$

In legacy network switch based systems, during expected normal use cases, the majority of the backup capacity resources allocated to a sector remain underutilized. The common base station design solution found in the majority of installations is to use additional routers and sector servers to increase the number of available sub-nets in the system. In low demand situations, these additional ports are fixed to the macrocell sector and are unused. These ports cannot be easily transferred to adjacent sectors or towers.

The underlying issue is that even with the use of millimeter backhaul the availability of routing ports in traditional network switches remains relatively static. New pathways to the centralized service center must be re-commissioned affecting the entire network. In most legacy systems prudent cost reductions require that only predicted network traffic load patterns are considered when designing the microcell distribution. If the base station is operating under stress, the result is an overall reduction in system performance. The solution lies in finding a methodology that both increases bandwidth and allows simple dynamic allocation and access to the increased bandwidth.

3 Edge Network Model

The most reliable 5G base station distribution model to date, which has been demonstrated in-situ, requires less than 1km between sector clusters. In most cases, urban models for high population centers suggest a 200-meter cell spacing. [3], [4]. The propagation models used in this proposal follow the (ETSI) European Standards Organization study [4] for an urban propagation model demonstrated in Equation 2 which models high density urban wireless networks. Highly dense urban models suffer from dense user traffic, propagation blocking obstacles, and multiple overlapping microcells. The maximum expected intra-base station spacing lies within the proposed 5km macrocell placement limit [4]. This limit also falls well below the threshold for significant optical dispersion interference between the closely packed optical channels required within the photonic switches. This suggested 5km limit also eliminates the need for optical amplifiers or repeaters between the microcells and the optical macrocell. Microcells are projected for 500m spacing however the ultra dense urban environment may make 200m more likely. The proposed networking topology model and router definitions used in this study reflect this limitation.

$$PL_{UMa-LOS} = \begin{cases} PL_1 & 10m \le d_{2D} \le d'_{BP} \\ PL_2 & d'_{BP} \le d_{2D} \le 5km \end{cases} \tag{2}$$

The test model follows the ITU specification [12], for channel allocation. The number of occupied channels in the fiber connection between stations contains a finite number of substreams covering N sector nodes in the system. This implies that there is a probability p that K active nodes are accessing the network at any given instance. In high traffic demand areas, this load changes dynamically and is expressed in Equation 3.

$$\binom{N}{K} p^K (1-p)^{N-K} \tag{3}$$

Consider that the majority of network traffic lies in simple data transfers such as email and web browsing requests, or similar nonstreaming operations. The emerging 5G network proposals offer much higher data rates to provide more advanced services. These services have near real-time prioritized packets, such as (VOIP) voice over IP and video streaming. In these use cases, the network path must stable to allow a steady non-blocking stream regardless of sector load. In most cases, the system designer builds the edge server to a predicted load model with approximately 20% overhead to handle additional network demands. Over-designing a sector results in increased power consumption, installation and maintenance costs for a given cell tower.

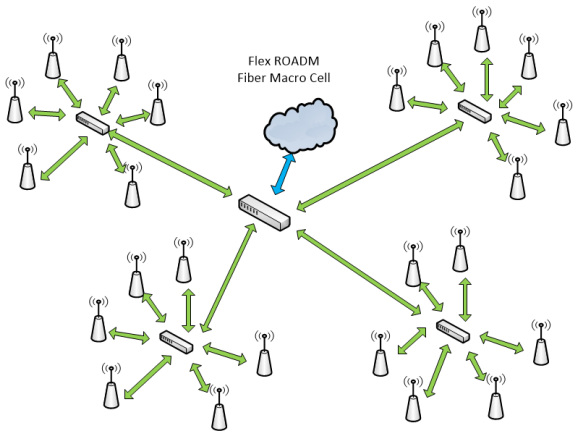


Figure 2: Flex Fiber ROADM Switched Backhaul Macrocell

Edge networking is a method that moves the control and access to the network directly to the end user node, at the edge. The optical network model proposed moves network access to the point of use. In traditional networks, multiple paths to a macrocell require fixed switch port allocations. In contrast, using the extreme bandwidth afforded by tunable laser transceivers and multiple wavelength selectable photonics switches allows the system manager to dynamically allocate bandwidth to meet the needs of the base station. When load demand increases, unused bandwidth can be allocated dynamically, without changing or adding existing equipment. Pathways to the Wide Area Network (WAN), are replaced with user selectable optical wavelength channels virtualizing port connections within the network. The photonic user reconfigurable switch

embedded in each NIC allows all microcells to be interconnected and any backhaul pathway on demand, as in Figure 2 using the proposed solution model.

In the proposed solution the edge network wireless interface and the available capacity of the sector radio modem are finite and represents the only limitation to the system architecture. Predicting 5G deployment is difficult, the base station location remains static, while the connection load and bandwidth demand are intrinsically dynamic. In switch based topologies, priority signals with high data rate connections experience interference with the routing requirements of new connection requests. The radio hand-off and other management traffic requirements use much smaller packets in the transport model and can be handled in a secondary network designated for command and control. In edge networking, parallel network pathways are common to alleviate traffic congestion. An increase in the number of users and therefore required connections sharing backhaul access pathways cut into available port bandwidth, even for simple management tasks. The result is an increase in asynchronous packet collisions on the central network access ports that are simultaneously attempting to stream high-speed data and handle the dynamically changing handoff management packets. The wide-ranging user access data rates necessary to provide efficient, reliable 5G level service requires a new methodology. Bandwidth allocation for the requested service and the number of network connections need to be either commissioned or placed back in the service pool on the fly without interfering with ongoing network traffic.

4 Plane Parallel Optical Network Topology

The concept of a "plane parallel" network is based on open and free connectivity between all attached nodes within a computer cluster topology. This concept can be easily adapted to data delivery services found in cell tower networks. The optical network topology further introduces the concept of virtualizable interconnects or ports made up of assigned combined optical and temporal data streams. This access flexibility enables the system administrator to create connections or ports on demand within the available optical bandwidth, which on a single fiber is currently in excess of 14Tb/s [13]. The underlying concept is to separate the optical wavelengths into service channels to parallelize the number of active networks within the system. Current research involves methods for managing high-speed data streams in optical systems. However, these methods are more of an adaptation of existing switch technology involving aggregated lower tiered switch networks. Once an optical pathway is set it remains relatively static to service the Internet backbone pathway, [14], [15].

The proposed optical backhaul uses extreme bandwidth networking interfaces to create a plane parallel network topology that allows multiple networks to reside on the same backbone without interference with adjacent traffic and increased delays. The proposed solution allows the sector control system to quickly establish and bring up additional sub-networks and assign service temporal data streams as required. The newly assigned virtualized interconnections allow system administrators to handle increased data loads and temporary high throughput services within the plane parallel topology. All access is direct to the optical network without reassigning existing pathways or switching through other physical ports as within a tiered switch system. Redundancy and resiliency are provided through access to multiple fibers. Optical networks have direct scalability intrinsic to this methodology through the use of redundant fiber bundles.

Networks and sub-clusters can be commissioned and decommissioned on the fly as opposed to required downtime through manipulation of the physical switch port and cable swaps between the installed macrocells. The basis for this plane parallel system is the use of photonic (ROADM) Re-configurable Optical Add Drop Multiplexer optical switch device and wavelength tunable transceiver lasers. These photonic interfaces are embedded directly within each network interface controller (NIC) located at the microcell sector interface. The central traffic management software can program any ROADM device to switch any frequency to a specific fiber. Using the ITU channel specification, the single fiber can handle 72 200Gb/s wavelengths simultaneously. The ROADM device allows one or more of these wavelengths to pass through reducing unnecessary traffic at the end node. A single fiber can easily handle multiple macro cells at the proposed full capacity of the 5G microcells. Legacy protocols are all supported through a required new socket layer control interface that can encapsulate existing protocols delivered to the NIC. Pre-assigned management and service data streams are located on separate wavelength networks within the optical interface. Service requests are handled as a task-oriented or process assignment which allows individual control of both the node service commissions and separation of parallel sub-networks providing requested services.

The centralized system scheduler-management application utilizes a common management network to respond to service requests from sector modems. Since all nodes are simultaneously available on the parallel fiber interface, a method is provided to handle handoff and assignment to active data streams that utilize a process ID addressing scheme rather than the common physical address used in existing network protocols. Service to the user is assigned a process ID, in email text legacy phone services these packets transmitted in the backhaul with the process or service connection ID and the IP addressing is decoded at the sector micro cell interface. If multiple users are streaming video this process ID can be parallelized to send the live video, or conference meeting stream to multiple users without establishing separate connections.

4.1 Optical Network Topology

The proposed network takes advantage of the extreme bandwidth availability using existing optical network interface equipment. A closed loop testbed was built to verify simulation models for both physical transmission of highly dense optical signals and the proposed transceiver modem used in the proposed NIC. Since this application is intended for use in a large scale distributed edge computing environment, one that can contain more than 10K connected nodes or devices, it is assumed that the maximum distance between any node is within a 1 km boundary to avoid dispersion effects found in long haul optical networks. In this configuration, any of the issues associated with optical dispersion and ICI (inter-channel interference) have negligible effects within this application.

The outstanding challenge involved with 5G backhaul bandwidth allocation is handling the dynamic increase in high-speed streaming services combined with the relatively short distance between handoff nodes. The number of users and the type of service requested is dynamically changing. It would be a waste of the proposed system's available bandwidth to simply assign optical channels of equal capacity to each microcell sector. The handoff from one radio to another is as simple as enabling an otherwise blocked temporal substream in the new radio already present on the optical interface, verifying the connection and dropping or blocking the stream on the previous radio. The proposed solution simulation example utilizes 50Gb/s tunable transceivers although 10Gb/s work just as well. The ROADM optical switches are embedded in each network interface card to allow a contention-less, colorless and bidirectional connection directly to one or more backbone fibers making up the plane parallel network distribution. Ideally, scalability of services requires adding fiber transceivers to the NIC interfaces in the microcell node using one or more of the dark fibers in the macrocell bundle. Telecom fiber installations typically include 12 to 24 fiber pairs for future use.

4.2 Flex Grid at the Edge

The use of ROADM based photonic switches flexible grid capability does not limit nor relegate the available channel spacing to one fixed size, rather the scheduler-manager application can assign varied channels boundaries within available requested optical range as needed to fill the optical boundaries. See Figure 3, which illustrates the flexible distribution of optical channels allocated. The impact of this optical boundary flexibility is that existing optical backhaul networks with fixed optical bandwidth boundaries can easily coexist with the addition of flexible channel allocation. Temporary ad-hoc networks commissioned for short term use can be combined into a larger channel if needed by a macrocell. Higher capacity channels can be assigned in a specific sector during peak usage hours and reassigned as the demand changes throughout the system.

The entire proposed network topology shares one or more common bidirectional fibers linking the sector modem at the edge thus bypassing the traditional distributed layered switch topology found in most 4G and legacy networks. The proposed method achieves a high-end user count with a flexible ratio of streaming (large packets) to service and management (small packet) distribution. Secondary transport methods evaluated varied between more common TCP/IP, and RDMA (Remote Direct Memory Access) favored by streaming clients. RDMA has significantly lower response latency because it systematically bypasses the processor core software by using the transport layer stack writing to memory directly. RDMA has an advantage in streaming services because large data files can be downloaded directly to the macrocell buffer to be distributed at a lower rate, freeing up the bandwidth to service other requests.

Proposed 5G market strategies favor streaming applications, and therefore the proposed solution will focus on the challenges of higher speed data and streaming packets. The optical router to radio interface will assume a 40Gb/s - 50Gb/s Ethernet connection assigned to one or more optical channels on a single fiber. The simulation of the control system assumes a minimum of four separate fibers available to each macrocell microcell cluster. Integrated multiplexing photonic switches allow for any four fibers to any four fibers multiplexing switch at the edge, controlled electronically by the sector node. The backhaul access can easily be scaled up to a 16-by-16 switch in future deployments. Similar networks have been proposed in HPC (high-performance computing) platforms [16, 13]. The embedded optical switch proposed allows the system to select one or more optical fiber pathways to adjacent towers or directly to the centralized backhaul Internet connection or data telecommunication center. Handoff communications are handled by the local towers rather than a centralized data center.

4.2.1 Available Optical Bandwidth Potential

The ROADM devices coupled with wavelength tunable lasers offer a "flexible" wavelength grid mapping capability that is completely user programmable. Acknowledging that this system proposed must exist in legacy networks, we will follow the ITU 694 wavelength channel assignment grid [12]. The ITU G694 specification currently has a defined maximum channel capacity of 100 GHz bandwidth each within the L band (1500 nm), made up of 72 optical channels of 100 GHz each from 190 THz to 194 THz [12]. Each 100 GHz channel has a standard capacity of 200 Gb/s each, using the common PAM (pulse amplitude modulation) 2 bit per Hz method. The data rates used in this study reflect this to illustrate the system viability using lower density modulation techniques. Emerging transceivers are demonstrating QAM 16, QAM 64 and OFDM over fiber, [13]. The term "flex grid" suggests that each available optical channel can be broken up into sub-channels as

long as the highest density channel boundaries remain static. Sub-optical channels are chosen such that the entire boundary is filled leaving no unused assigned space. A simple example is that of a single 100 GHz channel. The highest bandwidth under ITU, which can be subdivided into any combination equal to 100 GHz. Example, a single 50GHz channel, a single 25 GHz channel and the remaining can be a second 25 GHz channel or two 12.5 GHz channels as needed. The scheduler manager makes this allocation based on the type of service (data rate) and the number of expected users (nodes) on the microcell sector radio. The ROADM device is programmable electro-optically allowing these boundaries to be assigned. The optical wavelength and bandwidth representing the allocated sub-network channel to pass either upstream or downstream along the fiber. Figure 3, illustrates the existing boundaries within the ITU grid [12], and the flexible boundary assignments within each ROADM switch NIC device. It is important to note that ROADM devices have been demonstrated operating outside of the ITU assigned channel grid constraints and can have a completely user programmable boundary for channel allocation, [17]. In the test bed for legacy network compatibility the ITU grid boundary definitions are used.

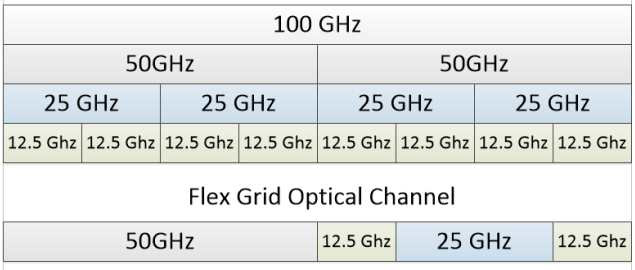


Figure 3: Single channel 100 GHz band divisions

Optical-Temporal data distribution is a method where the optical channels exist in a parallel plane. Each subsequent optical channel is broken up into temporal streams to accommodate varying bandwidth requirements filling otherwise empty space. The proposed solution provides that each assigned optical channel has the ability to further allocate the available bandwidth with a subset of temporal data-streams. Since most individual user connections cannot fully utilize 200 GB/s and promised 5G service is no more than 5 GB/s, the wide optical bands are filled with multiple data streams separated by time. A TDMA based set of temporal sub-data streams are included in the proposed solution shown in Figure 4. This hybrid optical channel/temporal data stream method allows the management controller to fully utilize the extreme bandwidth of the fiber, without using de-aggregation fixed port network switches. This arrangement is desired because the bulk of the connections will be asynchronous and flexible in size. Connection assignments are made using a single data stream or can be combined into compound streams when higher throughput is required by the service requested.

Table 1: Single fiber sub-channel data capacity

Optical Bandwidth	Modulation Density	Channels Available	TDMA Streams	Stream Bit Rate
100 GHz	200 Gb/s	72	2304	6.25 Gb/s
50 GHz	100 Gb/s	144	4068	3.125 Gb/s
25 GHz	50Gb/s	288	9216	1.56 Gb/s
12.5 Ghz	25 G/s	576	18432	781 Mb/s

Once a connection has established the wavelength and temporal stream assigned to the connection are given a unique connection ID code used as a compact routing header. The resulting connection assignment serves as a virtual port and is defined as a specific optical channel with one or more temporal data stream. The routing pathway is therefore further defined as Fiber 1...n, ITU channel wavelength 1...n, and temporal data stream 1 + n + m. One or more end user can share a single optical channel and temporal stream using the assigned connection ID code in the proposed socket header as a delivery method.

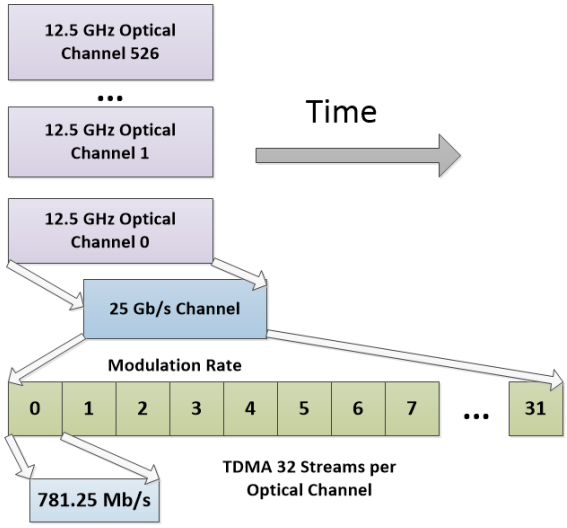


Figure 4: Parallel Optical Temporal Streams

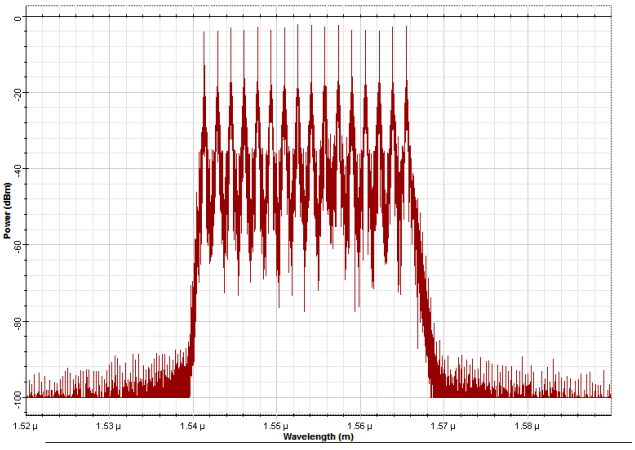


Figure 5: Parallel Optical Wavelengts

The sub-data streams are assigned by the scheduler-

manager software when a connection is commissioned for a specific service or process. A channel request packet is sent to the scheduler manager node for increasing or relinquishing assigned bandwidth back to a common resource allocation pool. The number of temporal sub-streams assigned is entirely up to the schedule manager. In the example application, each ITU designated optical channel has been assigned 32 temporal data slots within each optical transport band. A single base station can supply 9,216 1.56 GB/s data streams using a 50 Gb/s laser transceiver on a single fiber. Individual connection capacity is calculated using existing radios modulation capability coupled with the assigned bandwidth of the optical channel. In the test bed, example channel capacity is limited to a simple optical PAM modulation signal and using existing ITU defined channel boundaries. Simulation indicates that the multiple wavelengths can easily coexist at these channel boundaries. Figure 5, shows a simulation of 16 ITU sub channels all using 40 Gb/s lasers over a single 1 km fiber. A breakdown of an available channel using the suggested 32 sub-temporal data slots is shown in Table 1. This table contains the channel capacity for each of the assigned optical bands and temporal sub-channel count and the number of supported streams and capacity of each. A single fiber with 72 100 GHz optical channels can contain up to 14.4 TB/s capacity using this method. The backhaul quad fiber integrated edge switch has a capacity for 57.6 Tb/s.

The proposed photonic NIC contains packet buffers in the physical layer with intrinsic transceiver buffer load and unload times dependent on buffer size and memory access speed. The use of temporal data slots allows the system to synchronize the projected arrival time of the next data slot to match the physical load time of the buffers used. In most cases using temporal sub-channels allows for data buffers to be loaded or unloaded while waiting for the next available data slot, thus limiting transmission gaps within the optical channel.

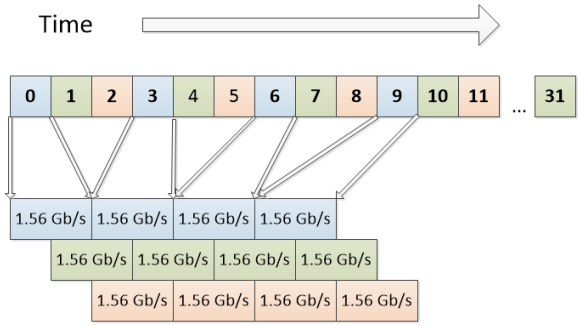


Figure 6: Aggregated staggered Temporal Data Stream

The chart in Figure 6 shows a typical scheduler assignment that uses four lower rate temporal data streams of 1.56 Gb/s each to be combined as a single 6.25 Gb/s virtualized port. In this example, the assigned multiple temporal slots are staggered to match the buffer fill times. The result is an effective zero buffer wait state for the physical transceiver buffers. In high bandwidth transmission non-aggregated packets that make up the bulk of networking traffic leave behind large gaps in the available transmit times. The temporarily assigned optical channel and sub-data stream act as virtualized connections at the socket layer. In the proposed network a commissioned modem tunes the laser and the ROADM receiver filter to a specific optical frequency and the electronics receiver uses the frame-sync timing signal from the command and control stream to synchronize the radio transceiver. The majority of packet buffering is handled directly at the edge as data is loaded into the physical layer buffer of the 5G radio modem.

4.2.2 Switch Scalability Model

The proposed 5G 5Gb/s use case can be realized using this distributed aggregation model using an optical transceiver with 25 Gb/s capacity shared across 32 temporal substreams. This transport data rate closely matches the proposed 5G maximum data rate and fits within the existing boundary of the existing ITU grid. As an example, in a single fiber network, consisting of eight sector 5G radio modems this represents non-blocking capacity for over 2000 individual 6 Gb/s data streams on a single fiber. A quad fiber ROADM switch increases this capacity per base station to 8000 non-interfering streams. Assuming the average expected end-user node typically requires around 1 Gb/s the number of available connections can be increased to 32K. The entire network capacity shares the plane parallel bandwidth assignment boundaries and an updated boundary map is distributed as a management packet on the control network. This allows all connected macrocells and microcells to know where a service stream is located within the optical/temporal channel assignment. Consider a 5G urban network with 6 Gb/s service capability and a projected average of 200 active connections at any one time per macrocell spaced at 1km. Using a shared fiber backbone connection 40 microcells spaced a maximum distance of 200 m from the macrocell can be serviced using only four shared fibers. The quad-optical switch proposed allows any sector radios access to the four backhaul fibers. The central scheduler manager handles the subnetwork assignments from the resource pool. When a connection is dropped, the channel allocation is demapped and returned to the resource pool. All connection requests use the single command and control optical network.

In this macrocell configuration, each of the four backbone fibers has a full spectrum carrying capacity of 14.5 Tb/s based on the ITU 100 GHz grid model. A single 16 sector microcell utilizing a quad fiber has direct access to 57.6 Tb/s of backhaul bandwidth. This translates to 9216 connection single temporal data streams each carrying 6.25 Gb/s covering the basic 6 Gb/s desired service with overhead for control messag-

ing, from table 1. If 20 Gb/s connections are required, there are two compromise solutions using the standard ITU grid. Three temporal data streams can be combined for 18.75 Gb/s on 3072 virtualized connections or four data streams can be combined for 2304 25Gb/s connections. The user programmable flexibility of the optical and temporal data assignments suggests that any combination up to the optical switch's maximum capacity for an assigned channel can be commissioned on the fly. The bandwidth allocation and connection commissioning are handled by the central scheduler manager node as required. Channel and sub-net assignments are negotiated by the requesting node for the type of service desired without impacting the other ongoing connections.

4.2.3 Test bed model

A testbed modem was used for the study using a two-channel transceiver with a functioning block diagram as shown in Figure 7 . The test bed consists of two (SPF+), Small Form-factor Pluggable, 10 G Tunable Lasers, and a pair of ROADM photonics devices. The testbed was used to verify that two adjacent dense packed optical channels can co-exist using a commercially available ROADM device [18], on the same fiber.

Measurements were made with adjacent lasers at full data rates and through separate photonics filters assigned before transmission. Inter-channel interference was measured on adjacent channels using a 20GS/s oscilloscope and found to be well below 100db specified by the ITU grid flex switch manufacturer, captured in Figure 8. Individual grid assigned optical wavelength channels can be duplicated as long as they are not located on the same fiber.

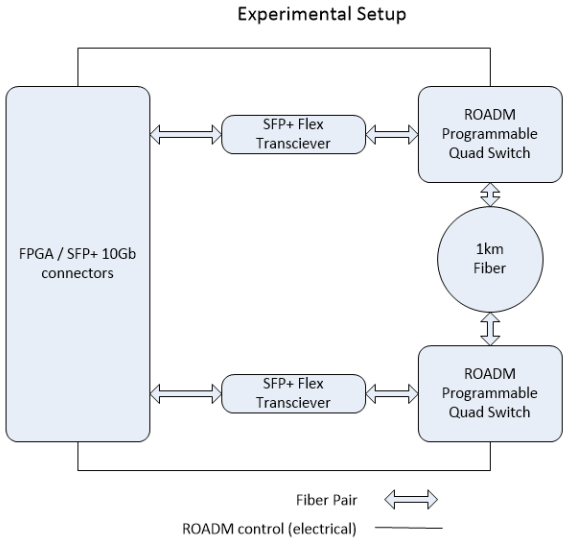


Figure 7: Experiment Schematic

Consider an extremely large control system, one defined as having more than 10K device connections. Distributive control is easily realized using this proposed topology. Lower bandwidth connections such as voice traffic, email or Internet browsing are assigned much smaller subdivided optical-temporal data streams. In most applications, these assigned data streams contain more capacity within the sub-channels then their existing counterpart high bandwidth switch based ports.

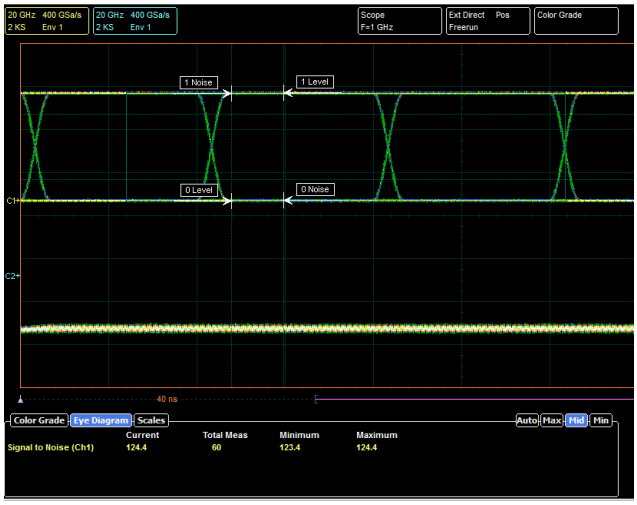


Figure 8: 10G received eye pattern with less than 100 db SNR

Advanced modulation techniques can increase the bandwidth of existing ITU channels or a custom optical channel distribution can be used in a homogeneous configuration. Specifically, one that does not have to conform to any industry-wide standards for channel spacing and center frequencies, such as a singularly connected private network. A recently demonstrated 400 Gb/s 64 QAM modulator [19], could increase the number of available 5 Gb/s links for a base station with a single fiber to just over 4500 service sub-nets, with a 500m separation limit. The bit rate limitation is due to the optical distortion effects common in higher order modulation methods.

4.2.4 Dynamic Port Assignments

A control access channel connection is maintained such that all available resources and radios can request service from the scheduler manager from a shared local network management channel. This same control path is shared by adjacent towers to facilitate handing off sub-nets instructing the new radio to switch the optical signal to the new sector. The process of bandwidth allocation and commissioning of system resources is independent of ongoing connections. In most applications, one or more optical-temporal streams are assigned as default command and control pathways for the desired service. The only one universally available command access network is permanently assigned to all macrocells and microcell NIC cards to facilitate a fallback command and management network path. During registration and commissioning the original transceiver MAC (media access controller) address, specific to the physical NIC in the sector radio, is used to identify the requesting device. Once commissioned all data transfers are sent with the process connection node ID code assigned during commissioning. The highest physical

layer of the optical modem selects the optical channel and captures the assigned temporal streams and checks each transfer with the process ID, to repackage and route the data to the end user. The control management system can directly access any assigned device by using the commissioned user node ID registered in the resource pool data base. During task and process communications, all network traffic after this uses the assigned optical temporal data paths and connection ID codes to distribute packets to their intended recipients.

4.2.5 Simulation Results

Performance of the proposed topology is limited to the latency in the photonics switching and modulation technique used. There are no multi-tier network switch buffers to fill, as are found in non-blocking switches. There is however a delay exists when an edge NIC is waiting for the next periodic arrival of the assigned temporal data stream.

During commissioning, the scheduler Manager uses the database of resources coupled with the requested data rate to match the temporal streams to the service NIC buffer fill and flush times. This temporal assignment mitigates extraneous delays and creates a near zero wait state delivery schedule. The total latency is defined by the stream availability delay and the electronic to optical conversion delay and modeled in Equation 4. Where total latency is the slot arrival time T_{slot}, added the data transfer propagation delay, $\frac{P_{length}}{D_{rate}}$ and finally electro-optical delay transforming the electronic signal to optical T_{e-o}. In the hybrid optical edge switch, there is no tiered switch induced latency. Ideally, the arrival time of the next available temporal data stream should match the time to fill the transceiver buffer. This method hands-off large stream buffering to the microcell sector radio rather than the shared network switching topology.

$$T_{total} = T_{slot} + \frac{P_{length}}{D_{rate}} + T_{e-o} \tag{4}$$

Throughput is dependent on a number of active links and their respectively assigned optical bandwidths. It is therefore defined by the average of the sum of the through-puts of all the subchannel links N_{link} as in Equation 5, where T_{e-o} represents the electronic to optical conversion delay.

$$Throughput_{avg} = \frac{\sum_{i=1}^{n} Throughput_{perlink} * i}{N_{link}} \tag{5}$$

In this study, the single pathway latency for a delivered packet was both simulated and measured in the test bed to determine the average delivery latency. The size of the packet used was a range of test packet sizes from 16 bytes (management request or acknowledge) and large scale streaming 16K packets and found to center around 4 uS in the 40Gb transceiver, and 3.5 uS in the 50Gb/s Ethernet over Infiniband transceiver as shown in Figure 9. The simulation used NS-3 with a custom model generated using the measured latency from the testbed, There is a slight increase in latency as the packet size is maximized due to the last point buffer fill times piling up during transmission. In most cases the entire receive buffer is available long before the next packet arrives when simulating a 6 Gb/s 5G user link.

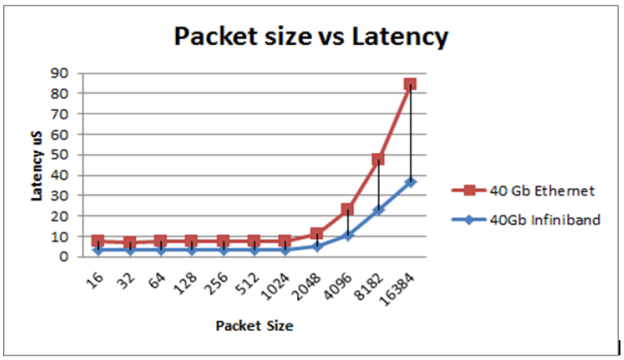


Figure 9: Transport latency vs. packet size

As the network load approached the maximum throughput of the assigned channels, the latency increased, primarily due to occasional loss of synchronization with the assigned data stream. This can be remedied by either increasing the staggered delay assigned slot or the size of the fill buffer in the transmitter-receiver interface. The lower network loads from 1 to 85% capacity remain reasonably static as shown in Figure 10. The temporal data streams are continuous, so it follows that the increase in latency at the higher packet sizes is due to the encapsulating of the segmented packets within the new socket structure. This effect was caused by the legacy TCP/IP protocols, including Ethernet ARP (address resolution protocol) included within the network stack. The optical transport header making up the outside layer using the process ID code to identify the packet at the destination, replacing the physical addressing used in the legacy protocols. Jumbo frames were not used in the study, so the limitation of 1500 bytes [20], remained requiring the transmission of several smaller packets when larger block sizes where transmitted.

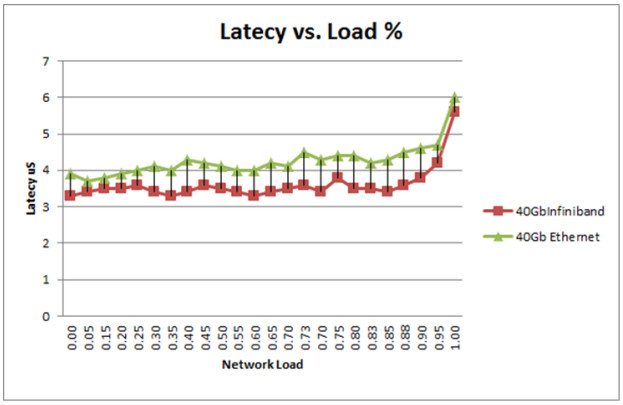


Figure 10: Transport latency vs. network load

5 Conclusion

The purpose of this study was to explore an alternative approach to the issue of bandwidth availability in 5G macrocells. The proposed approach was to use emerging technology for terabit networking and apply it as a (FTTC) fiber to the curb solution. There are three key issues addressed.

- Reconfiguration of the bandwidth to match dynamic loads.
- Reduce system costs while providing scalability.
- Provide a method for de-aggreagating 200 Gb/s streams directly to 6Gb/s streams without tiered switch network.

Simulations followed by testbed experiments show that the emerging ROADM photonics technology can be adapted to allow the network designer to provide a complete software-defined distribution network that is dynamically reconfigurable on the fly. The existing commercial hardware meets this constraint. Direct optical access requires replacing the Ethernet physical layer in the TCP/IP stack with a new layer that provides a bridge between the terabit optical interface and the 5G microcell sector modem 10G Ethernet NIC. Scalability requires a swap out of the SFP photonics transceiver device in the NIC to 40Gb/s model to support the expected 20Gb/s 5G service model. A prototype interface was tested and acted as a proof of concept for this work, which met the requirements of the 6Gb/s service model using a ROADM module [18] and an SFP+ 10 Gb transceiver through 500m of fiber. Certainly, within the confines of existing legacy optical network standards [12], it has been shown that it is possible to provide adequate 6Gb/s data streams co-existing on a single fiber without blocking or interference with legacy 4G and 3G traffic. Additional efforts are needed to fully refine and test a new physical (SDN) software-defined networking protocol layer to be applied to a fully expanded testbed that models an entire 5G macrocell. Future efforts will focus on the command and management protocol necessary for full deployment of this concept.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to thank Optiwave Inc, and Infinera Inc. for simulation software and flexgrid photonics technical information.

References

[1] T.S. Rappaport, S. Hu, R. Mayzus et al.,*Millimeter Wave Mobile Communications for 5G Cellular. It will Work, IEEE Access Vol1, pp 335-349, May 2013.*

[2] N. Bitar, *SDN and potential applicability to access net- works*, in Optical Fiber Communication Conf. and Exhibition OFC 2014, paper Th1G.4.piling

[3] M. Coldry, J-E Berg, L. Manholm, C. Larsson, J. Hasryd, *Non Line of Sight small cell backhauling sing microwave technology, IEEE Wireless Commun. Mag , Vol 51 No. 9, pp. 78-84, 2013.*

[4] ETSI. New ETSI Group on Millimetre Wave Transmission Starts Work. Technical report.

[5] C. Dehos, J.L. Gonzales. A. D. Dmoenico, D. Ktennas L.Dussopt, *Millimeter Wave Access and Backhauling: The solution to the Exponential Data Traffic Increase in 5G Mobile Communications Systems, IEEE Trans Wireless Commun. Mag , Vol 52 No. 9, pp. 88-95, Sep 2014.*

[6] X. Ge, S. Tu, G. Mao, C.X. Wang, T. Han,*5G Ultra-Dense Cellular Networks, IEEE Wireless Communications, 2016.*

[7] S. Hur, T. Kim, D.J. Love, J.V. Krogmeier, T.A. Thomas, A. Ghosh, *Millimeter Wave Access and Beamforming for Wireless Backhaul and Access in Small Cellular Networks, IEEE Trans Wireless Communications, Vol 61 No. 10, pp. 4391-4403, Oct 2013.*

[8] CISCO. Global Visual Networking Index: Globile mobile data traffic report forcast update, 2015-2020 (2015) White Paper

[9] S.F. Yumas, M. Valkama, J. Niemela, ,*Spectral and Energy Efficiency of Ultra Dense Networks under Different Deploymenmt Strategies, IEEE Commun. Mag , Vol 53 No. 1, pp. 90-100, 2015.*

[10] Intel Case study: Rethinking small cell business model(2012), Intel White paper

[11] J. Robson. Small cell backhaul: an analysis of small cell backhaul requirements and comparison of solution, Cambridge Broadband Networks Limited(2012) White Paper

[12] International Telecommunications Union *Spectral grids for WDM applications: DWDM frequency grid*, REC G.694.1 201202 I 2012

[13] N. Terzenidis, P. Maniotis, and N. Pleros, *Bringing OptoBoards to HPC- scale environments: An OptoHPC simulation engine, in Proc. 1st Int. Workshop Adv. Interconnect Solutions Technol. Emerging Comput. Syst., 2016, Art. no. 6.*

[14] X. Li, K. Kanonakis, N. Cvijetic, A. Tanaka, C. Qiao, and T. Wang, *Joint bandwidth provisioning and cache management for video distribution in software-defined passive optical networks, in Optical Fiber Communication Conf. and Exhibition (OFC), San Francisco, CA, Mar. 2014, paper Tu3F.5.*

[15] V. Lpez, O. Gonzlez de Dios, L. M. Contreras, J. Foster, H. Silva, L. Blair, J. Marsella, T. Szyrkowiec, A. Autenrieth, C. Liou, A. Sasdasivarao, S. Syed, J. Sun, B. Rao, F. Zhang, and J. P. Fernndez-Palacios, *Demonstration of SDN orches- tration in optical multi-vendor scenarios, in Optical Fiber Communication Conf. and Exhibition (OFC), Los Angeles, CA, 2015.*

[16] Y. Matsuoka et al. *20-Gb/s/ch high-speed low-power 1-Tb/s multilayer optical printed circuit board with lens-integrated optical devices and CMOS IC, IEEE Photon. Technol. Lett., vol. 23, no. 18, pp. 13521354, Sep. 2011.*

[17] Infinera The Next Generation of Coherent Optical, Advance Coherent toolkit for Super Channels (2016), Infinera White paper

[18] SANTEC 4 Channel Mux / Demux Module MDM-15 (2018), SANTEC data sheet

[19] G. Wei-Lu T. Sakamoto, T. Kawanishi, *Flexible high order QAM transmitter using tandem IQ modulators for generating 16/32/36/64-QAM with balanced complexity in electronics and optics, OSA Optics express Vol 21, No 5, pp. 6213-6223 Nov 2013.*

[20] M. Mathis, J. Semke, J. Mahdavi, T Ott *The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm CCR ACM SIGCOM (1997) vol 27 No. 3 ISSN 0146-4833*

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 70-77 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A Spectrum Sharing based Metering Infrastructure for Smart Grid Utilizing LTE and WiFi

Imtiaz Parvez, Arif I. Sarwat *

Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174

ARTICLEINFO

Article history:
Received: 12 December, 2018
Accepted: 09 February, 2019
Online: 15 March, 2019

Keywords:
3.5 GHz
AMI
cognitive radio
LTE-U
CBRS
spectrum sharing
Internet of Things (IoT)
smart meter
licensed assisted access (LAA)
WiFi

ABSTRACT

In advanced metering infrastructure (AMI) of smart grid, WiFi is an appropriate choice for its bidirectional communication requirement to transmit data to the billing center. But, WiFi functions in the free spectrum bands and LTE also requires to use the same free bands for its network expansion being licensed spectrum is limited and expansive. LTE and WiFi can operate simultaneously in the 3.5 GHz band (also known as citizen broadband radio service (CBRS)), which has large amount of free and clean spectrum. In this paper, we propose a smart grid metering infrastructure based on fixed duty cycled LTE and WiFi, where smart meters and its' data collectors (known as Access Point) use WiFi and LTE, respectively, for transferring data. Under a system level simulation environment, we investigated the LTE-WiFi coexistence performance in CBRS band considering a time division duplexing (TDD)-LTE associated with FTP traffic, and IEEE 802.11n (WiFi). The simulation performance demonstrates a good neighborhood coexistence between WiFi and LTE, which makes it a potential communication solution for the AMI.

1 Introduction

Smart grid is the advanced power and energy system that has been transformed from unidirectional power flow to bidirectional power flow. Moreover, it is employed with information communication technology among its entities for electricity supply to the consumers with reinforced control and efficiency [1, 2, 3, 4]. Advanced metering infrastructure (AMI) is an critical building block of smart grid as it builds communication bridge between metering data management service (MDMS) and consumer meters for consumption data transfer utilizing wireless networks [5, 6, 7, 8, 9]. The prominent communication standards for AMI are Zigbee, and WiFi, that use public frequency bands e.g. 5 GHz, 2.4 GHz, and 900 MHz [10]. Since unlicensed/public frequency bands will be shared, smart meters may need to coexist along with atypical techniques e.g. ZigBee and LTE in the same bands.

LTE is the long range broadband communication for exchanging voice and data [1, 11]. For the advancement of requirement, LTE requires to accommodate machine-to-machine (M2M) communication in addition to voice communication. Moreover, to satisfy the exponential increase of throughput requirement in LTE, spectrum shortage is a critical obstacle. In this regard, spectrum sharing among different wireless technologies could be a promising solution. However, this sharing approach has its own implementation hurdles. Additionally, public (license free) spectrum can be used in conjunction with licensed spectrum. In this regard, 3GPP working group is studying on the license assisted access (LAA) of LTE in the free bands [12].

WiFi[1] is a prominent short range communication protocol which utilizes a distributed coordination function (DCF). Its channel access mechanism performs four-way handshaking and carrier sensing [13]. The WiFi DCF mode utilizes clear channel assessment (CCA) technique for the packet transmission. The CCA includes energy detection and carrier sensing mechanism to detect the state of channel- whether it is in operation or not. WiFi node will cease transmission attempt for a random time period if the interference level crosses CCA threshold. This back-off method avoids packet collision that may happen due to coexisted LTE network transmission.

*Corresponding Author: Arif I. Sarwat, Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174, Email: asarwat@fiu.edu

[1] Unless or otherwise saying, IEEE 802.11n version will considered as WiFi in our study.

https://dx.doi.org/10.25046/aj040209

In contrast, the LTE technology is comparatively flexible and systematic. LTE utilities dynamic scheduling for its users. The main obstacle for coexisting WiFi and LTE system in the identical band is the data transmission technique. WiFi uses CSMA/CA protocol for the transmission of OFDM. On the other hand, LTE uses the dynamic scheduling in OFDM access through which data is transmitted to several UEs simultaneously at low rate with proper time and frequency allotment [14]. LTE reserves channels to make transmission simultaneously. On the other hand, WiFi implements carrier sensing before the packet transmission. Therefore, LTE transmission will block the WiFi transmission most likely in the coexistence scenario.

Recently, 3.5 GHz band (also known as citizen broadband radio service (CBRS)) has been released for public use, which is to be shared [15, 16]. According to the guideline, the users can be classified into three classes: first tier users, second tier users, and third tier (general) users. In general, third tier/general users access the CBRS spectrum, giving priority to the first and second tier users. In several cases, third/general users can use full 150 MHz of bandwidth in the absence of first and second tier users' activity [17]. At the worst case scenario, 80 MHz spectrum will always be available for third tier/general users when second tier users are active and operated outside of first tier users' zones. This large amount of spectrum can provide clean channels for various wireless communication applications such as smart grid metering data communication [18, 19, 20, 21]. WiFi and LTE use the CBRS band as the potential general (third tier) users in our study.

In this study, we expand our previous work in [22], which introduces a AMI architecture based on WiFi and LTE coexistence. In the architecture, WiFi is used in smart meters for transferring data to Access point (AP). After collecting data from a group of meters, AP transfers the data to MDMS utilizing the LTE. In our framework, we consider an integrated LTE-WiFi system where LTE BS and WiFi APs are connected through IP layer. Following this, we investigate the performance of LTE-WiFi coexistence in the CBRS band considering both conventional personal mobile communication and AMI communication. For system level simulation, a time division duplexing (TDD)-LTE and WiFi are considered in a seven cell hexagonal layout. LTE uses a fixed duty cycle of a transmission period for its transmission, and WiFi transmits in the remaining period, in contrast. The simulation performance exhibits a harmonious coexistence relationship between WiFi and LTE. Since CRBS posses a huge chunk of free and clean spectrum, AMI based on LTE-WiFi coexistence operating in CBRS can be a potential communication resolution for smart grid. The contributions of our work is as follows:

1) We introduce a smart grid metering infrastructure based on fixed duty cycled LTE and WiFi for the first time, where LTE and WiFi shares the same spectrum band.

2) Our proposed spectrum sharing method ensures good neighborhood spectrum sharing with an option of adjusting duty of LTE transmission.

3) Our spectrum sharing technique enhances the spectral efficiency significantly.

4) We propose the usage of recently release CRBS band for metering infrastructure which can provide large amount of free and clean spectrum.

The subsequent sections are arranged as follows. The literature review on LTE-WiFi coexistence and AMI communication is discussed in Section II. The coexisted system model of LTE-WiFi coexistence in 3.5 GHz band is illustrated in Section III. Deployment scenario and performance results are illustrated in Section IV. Lastly, Section V summarizes the whole work. Some of the acronyms used in this paper are presented in Table 1.

Table 1: List of Acronyms.

3GPP	3rd Generation Partnership
AMI	Advanced Metering Infrastructure
AP	Access Point
APP	application
CSMA/CA	Collision Sensed Multiple Access/Collision Avoidance
CCA	Clear Channel Assessment
CBRS	Citizen Broadband Radio Service
DCF	Decentralized Frequency Control
EDCA	Enhanced Distributed Channel Access
EPC	Evolved packet core
EPC	Enhanced Packet Core
FCC	Federal Communications Commission
FBE	Frame Based Equipment
GTP	GPRS tunneling protocol
PAL	Prioritized Access License
PHY	physical
PDPC	packet data convergence Protocol
GAA	General Authorized Access
IP	Internet Layer
LAA	Licensed Assisted Access
LBT	Listen Before Talk
LBE	Load Based Equipment
LLC	logic link control
LTE	Long Term Evolution
M2M	Machine-to-Machine
MDMS	Meter Data management Service
MAC	medium access control
OFDM	Orthogonal Frequency Division Multiplexing
PPDU	PLCP Protocol Data Unit
PL	Path Loss
RLC	radio link control
SINR	Signal-to-Interference-plus-Noise Ratio
TPC	Transmission Power Control
TTI	Transmission Time Interval
UE	User Equipment
UDP	user datagram protocol

2 Literature Review

The variants of LTE working in the public/free bands can be categorized into two groups: (1) LTE-U and (2) LTE-LAA [23]. LTE-U was developed by industry consortium [24]. It uses simple mechanism and excludes modification in the air interface structure of LTE system. It is founded on the LTE release 10-12 aggregation protocol and does not embrace LBT [25]. On the other hand, LTE-LAA is based on 3rd Generation

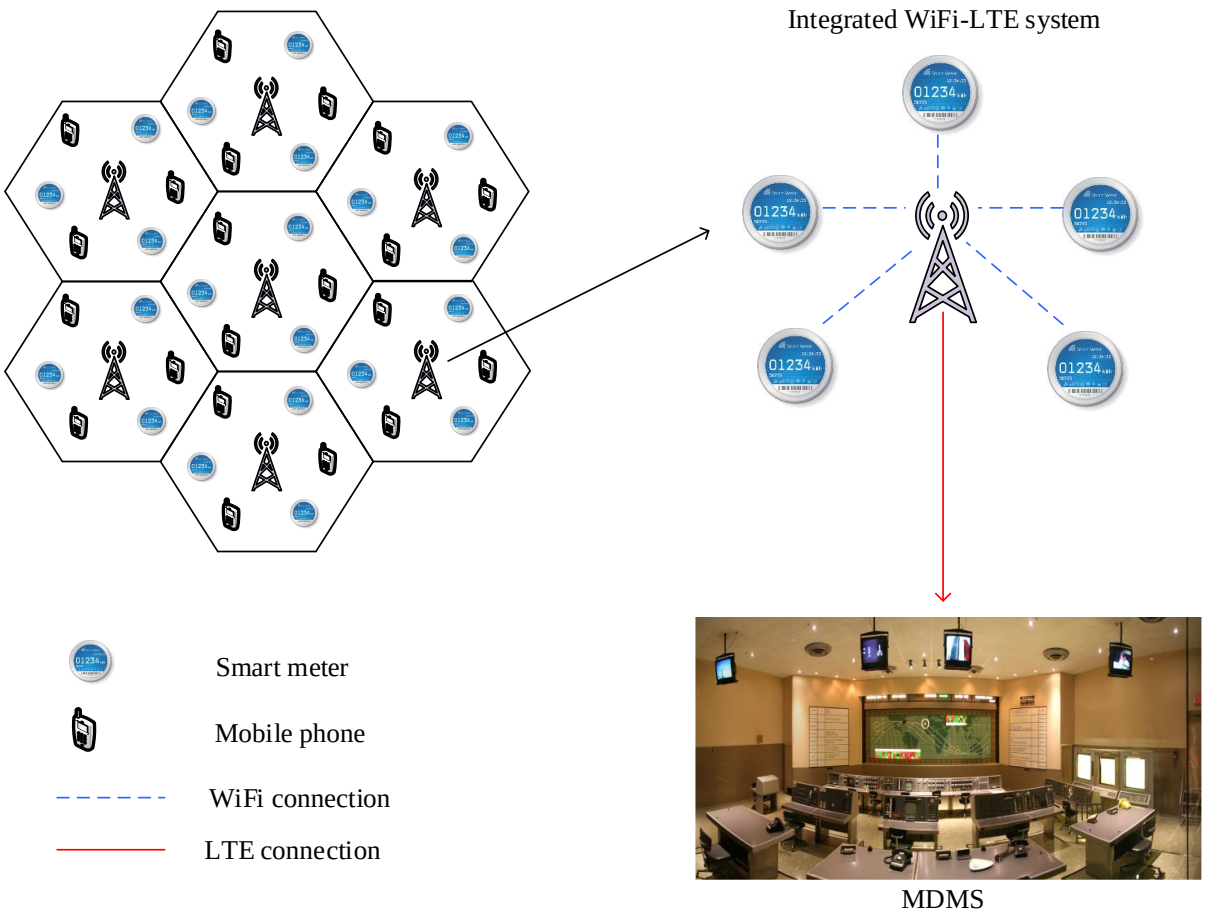


Figure 1: Architecture of smart grid metering infrastructure using LTE and WiFi on a collocated cell layout.

Partnership (3GPP) Release 13, which aims to develop a single global framework [26, 27].

In the literature, mainly three techniques have been proposed for coexistence between WiFi and LTE-U/LAA. They are- 1) listen before talk (LBT); 2) transmission gap; 3) dynamic channel selection. In [28], least congested channel search and adaptation of channel bandwidth are proposed for LTE[2]. Qualcomm proposed an interference level based effective channel selection technique in [29]. If the interference at the operating channel crosses the threshold value, LTE alters the channel with interference measurement before and during operation at both the network and equipment side. In Japan and Europe, LBT is compulsory for data transmission in the unlicensed band. The LBT techniques can be divided into two groups- frame based equipment (FBE) and load based equipment (LBE). In FBE based LBT, a fixed slot of frame is reserved for transmission where CCA is performed [30, 31]. If the channel is empty, the transmission is attempted. Otherwise, it will wait for the next frame. On the other hand, LBE based LBT is demand driven and the user equipment finds a clear channel for transmission [32, 33]. It performs extended CCA (ECCA) for clear channel access. Carrier aggression from licensed to public band is introduced in [34] using clear-to-send (CTS) and request-to-send (RTS) together with LBT. In [35], a technique of blank subframe allocation is introduced in LTE subframe, in which WiFi transmits. In [36], a identical approach is proposed, where n out of 5 sub-frames of LTE is reserved for the transmission of WiFi.

In [13], the coexistence performance of hot-spot indoor scenario is explored using a semi-static system level simulator. The study found that WiFi's performance deteriorated more significantly than the performance of LTE when operated in the same band. In [37], the similar result has been found for coexistence system of ZigBee and LTE, where the performance of ZigBee is affected more compared to that of LTE. [38] explored the usage of different communication networks and recommended to use LTE for low density scenarios (i.e. rural regions) and WiFi for high density scenarios (i.e. urban regions). Meter data communication using the hybrid WiFi/LTE configuration is introduced in [39], where LTE is kept on the upper layer and WiFi in the bottom layer. However, LTE and WiFi uses different spectrum bands in this architecture and there is no spectrum sharing aspect in this study.

In our study, we introduce a fixed duty cycle based coexistence for AMI of smart grid, where LTE and WiFi shares the same spectrum band. Additionally, we consider an integrated LTE-WiFi system where LTE BS and WiFi APs are connected through IP layer. WiFi is

[2]Unless and otherwise specified, LTE will be considered as LTE-U or LTE-LAA throughout this study.

used for meter-to-meter and meter-to-AP data communication. On the other hand, AP uses LTE to transfer data to MDMS. The duty of LTE transmission can be adjusted based on the data amount.

3 System Architecture

Let us assume, a coexisted network architecture consists of WiFi and LTE (LAA/LTE-U) operating in CBRS, as shown in the Fig. 1. Smart meters utilize WiFi and APs utilitie LTE for data transfer, in contrast. In addition, LTE BS and WiFi AP are attached together in the collocated environment. WiFi APs collect smart meters' data and forward them to the interconnected LTE BS. Afterwards, LTE BSs transfer data to MDMS through long range communication. Fig. 2 illustrates the protocol mapping of different components of LTE network and WiFi system. The PHY layers of WiFi AP and smart meters are connected together through wireless channel. Additionally, the IP layers of LTE BS and WiFi AP are integrated together in our proposed configuration. The data exchange among enhanced packet core (EPC), LTE BS, and MDMS are carried out according to standard LTE system [1].

We assume, the sets of LTE BS, WiFi STAs (i.e. smart meter), WiFi APs (i.e. collector of data from meters), and LTE UE (i.e. MDMS and other UEs) are marked as S_l, U_w^i, S_w, and U_l^j, respectively. Besides, LTE BS j, LTE UE/MDMS m, WiFi AP i, and meter/WiFi STA l transmission power are denoted by p_r^j, p_r^m, p_r^i, and p_r^l.

The channel gain values from LTE UE a to WiFi AP j, from WiFi STA/meter x to WiFi AP j, from LTE BS $b(i \neq b)$ to WiFi j and, from LTE BS i to WiFi AP j are $h_{j,r}^a$, $h_{j,r}^x$, $h_{j,r}^b$, and $h_{j,r}^i$ respectively.

During the data reception, the signal to interference plus noise ratio (SINR) of WiFi AP j from meter/WiFi STA x at the r the resource block [40] is

$$\mathrm{SINR}_{j,r}^x = \frac{h_{j,r}^x p_r^j}{\sum h_{j,r}^a p_r^a + \sum h_{j,r}^i p_r^i + \sum h_{j,r}^b p_r^b + \sigma^2}, \quad (1)$$

where σ^2 is the noise variance. A low SINR results poor throughput whereas high SINR ensures good throughput.

The received bit N_{B}^x at WiFi AP j from WiFi STA x [40] is given by

$$N_{\mathrm{B}}^x = \mathrm{BT} \sum \log_2(1 + \mathrm{SINR}_{j,r}^x), \quad (2)$$

where B and T (T=$\sum r$) are the bandwidth and transmission time, respectively. The received bit number is dependent on the SINR value.

The throughput of WiFi STA/meter x during the up link (UL) can be expressed [40] as

$$C^x = \frac{N_{\mathrm{B}}^x}{T_{\mathrm{tx}} + T_{\mathrm{wait}}}, \quad (3)$$

where T_{wait} and T_{tx} represent the wait time and transmission time of WiFi, respectively.

For down-link capacity calculation, similar equations: (1)-(3) are applicable.

The arrival rate of traffic for both WiFi and LTE is λ. The function relating delay of incoming packets (d) [40] is then

$$f(d) = \lambda e^{\lambda d}. \quad (4)$$

4 Deployment Scenario and Simulation Results

As illustrated in Fig. 1, a coexisted network layout of 7 cells is considered to investigate the system performance. A Matlab simulator founded on 3GPP standard was used for simulation similar to [13, 41]. For each integrated WiFi AP and LTE BS, 10 LTE UEs and 10 smart meters (WiFi STAs) are dropped randomly in each cell. One of the 10 LTE UEs is used as the MDMS. For both WiFi and LTE, the data arrival rate is kept same as $\lambda_{\mathrm{WiFi}} = \lambda_{\mathrm{LTE}} = 2.5$ packets/second. The PHY and MAC layer of IEEE 802.11n and LTE are enforced in the simulation scenario. Single UE is scheduled for DL/UL during a transmission time interval (TTI) and the corresponding SINR is sent to the BS. During one subframe of transmission, bandwidth is divided among all UEs based on request and waiting LTE UEs. TABLE 2 summarizes the simulation parameter for LTE, where values were chosen according to 3GPP LTE standard [12].

Enhanced distributed channel access (EDCA) and advanced clear channel assessment (CCA) have been enforced for WiFi channel access mechanism i.e. CSMA/CA. After receiving a beacon signal, all WiFi STAs (i.e. meters) with traffic will be in competition for accessing channel. Data transmission or reception will be at postponed without receiving a beacon signal.

Table 2: PHY and MAC Layer Parameters for LTE.

Parameter	Value
Frequency band	3.5 GHz
Bandwidth	20 MHz
Transmission power of DL transmission	15 dBm
Velocity of UEs	0 ms
Transmission power of UL transmission	PL Based TPC
Frame duration	10 ms
Type of scheduling	Round Robin
P_0	-106 dBm
TTI	1 ms
Packet arrival rate (λ)	2.5

The WiFi STA will sense for a free channel before any kind of transmission. Transmission will take place only if the channel is in idle, otherwise it will back off. After a randomly chosen back off time, next transmission

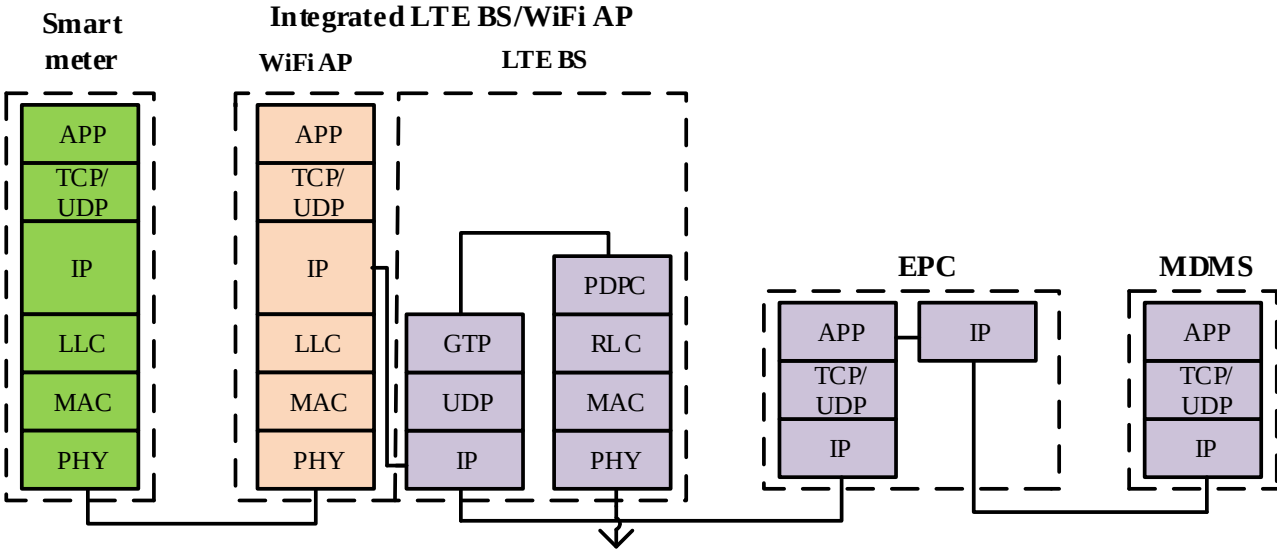


Figure 2: Protocol mapping among various entities of LTE and WiFi system.

will be attempted. TABLE 3 summarizes the WiFi simulation parameter used in the simulation [13, 18, 42].

The abstract of PHY layer is used for calculating Shannon capacity of LTE and WiFi at the $4\mu s$ granularity of WiFi OFDM symbol period. FTP traffic model-2 is applied for both LTE and WiFi traffic [43]. In this study, duty cycles- 20%, 40%, 60% and 80% of 50 ms time period are utilized for LTE transmission and the rest of 50 ms, i.e. 80%, 60%, 40% and 20% are used for WiFi transmission, respectively. The data rate performance of coexisting WiFi and LTE system is presented in Table 4. For 20% duty cycle, the LTE throughput is 10.3 Mbps and the throughput of WiFi is 155.2 Mpbs.

Table 3: PHY and MAC Layer Parameters for WiFi.

Parameter	Value
Frequency band	3.5 GHz
Bandwidth	20 MHz
Transmission power of Downlink/Uplink	23 dBm
Velocity of STA/meter	0 ms
Category of access	Best Effort
Protocol for MAC layer	EDCA
Sensing threshold of CCA	-82 dBm
Energy detection threshold of CCA	-65 dBm
Number of PPDU service bits	16 bits
Number of PPDU tail bits	12 bits
Window size for contention	$\mathcal{U}(0,31)$
Noise figure	6
Interval for beacon transmission	100 ms
Threshold of symbol detection in OFDM	10 dB
Threshold of beacon error ratio	15
Arrival rate of packets (λ)	2.5

For 40% duty cycle of LTE, the throughput of LTE and WiFi are 18.8 Mbps and 111.78 Mbps, respectively. For 60% duty cycle of LTE, the throughput of LTE and WiFi are 36.3 Mbps and 36.1 Mpbs, respectively. The throughput of LTE is boosted to 38.6 Mbps after increasing the duty cycle of LTE to 80%. However, WiFi capacity is reduced to 31.2 Mbps. Therefore, for increment of LTE transmission duty cycle, the LTE capacity is improved and WiFi is degraded drastically. The reason behind the WiFi throughput degradation is the increased transmission back off on the extended period of LTE transmission.

The energy efficiency (EE) performance of coexisted systems is demonstrated in Table 5. It is noted that the EE of LTE is improved with the increment of duty cycle of LTE. The EEs of LTE at 20% and 80% duty cycle are 3.32×10^8 bits/joule and 1.245×10^9 bits/joule, respectively. On the other hand, the EE of WiFi is degraded with the increase of LTE duty cycle. The EEs of WiFi at 20% and 80% duty cycle of LTE are 7.76×10^8 bits/joule and 1.56×10^8 bits/joule, respectively. More significantly, the overall EE of the coexisted system continues to improve with the increment of LTE transmission duty cycle. The overall EE is boosted from 1.008×10^9 bits/joule to 1.401×10^9 bits/joule. This reflects a good neighborhood relationship between LTE and WiFi regardless of degradation of overall throughput of the coexisted system.

The SINR distribution of coexisting LTE and WiFi system is illustrated in Fig. 3. For 20% and 40% duty cycle of LTE transmission, WiFi has better SINR distribution over LTE system. This is reflected in Fig. 3(a) and Fig. 3(b). For the increment of LTE duty cycle to 60% and 80%, the SINR of LTE system improves while the SINR of WiFi degrades consequently. This is demonstrated in Fig. 3(c) and Fig. 3(d), respectively.

In urban or suburban areas, large number of smart meters will use WiFi for sending consumption data to AP, and later the collected data will be sent to MDMS using LTE. Therefore, more opportunity of accessing channel by WiFi is desirable in this case. In this regard, 20% and 40% duty cycle of LTE transmission can be prudent choice for AMI infrastructure. On the other

Table 4: Capacity of the coexisted LTE-WiFi system

Duty cycle	LTE			WiFi		
	Down link (bits/second)	Up link (bits/second)	Total (bits/second)	Down link (bits/second)	Up link (bits/second)	Total (bits/second)
20%	9.15×10^{6}	1.153×10^{5}	1.030×10^{7}	8.343×10^{7}	7.176×10^{7}	1.552×10^{8}
40%	1.667×10^{7}	2.17×10^{6}	1.884×10^{7}	6.435×10^{7}	4.742×10^{7}	1.1178×10^{8}
60%	2.81×10^{7}	8.16×10^{6}	3.63×10^{7}	2.45×10^{7}	1.16×10^{7}	3.61×10^{7}
80%	2.71×10^{7}	1.15×10^{7}	3.86×10^{7}	1.66×10^{7}	1.46×10^{7}	3.12×10^{7}

Table 5: Energy efficiency performance of coexisted LTE-WiFi system

Duty cycle	LTE (bits/joule)	WiFi (bits/joule)	Total (bits/joule)
20%	3.32×10^{8}	7.76×10^{8}	1.008×10^{9}
40%	6.07×10^{8}	5.58×10^{8}	1.17×10^{9}
60%	1.171×10^{9}	1.80×10^{8}	1.351×10^{9}
80%	1.245×10^{9}	1.56×10^{8}	1.401×10^{9}

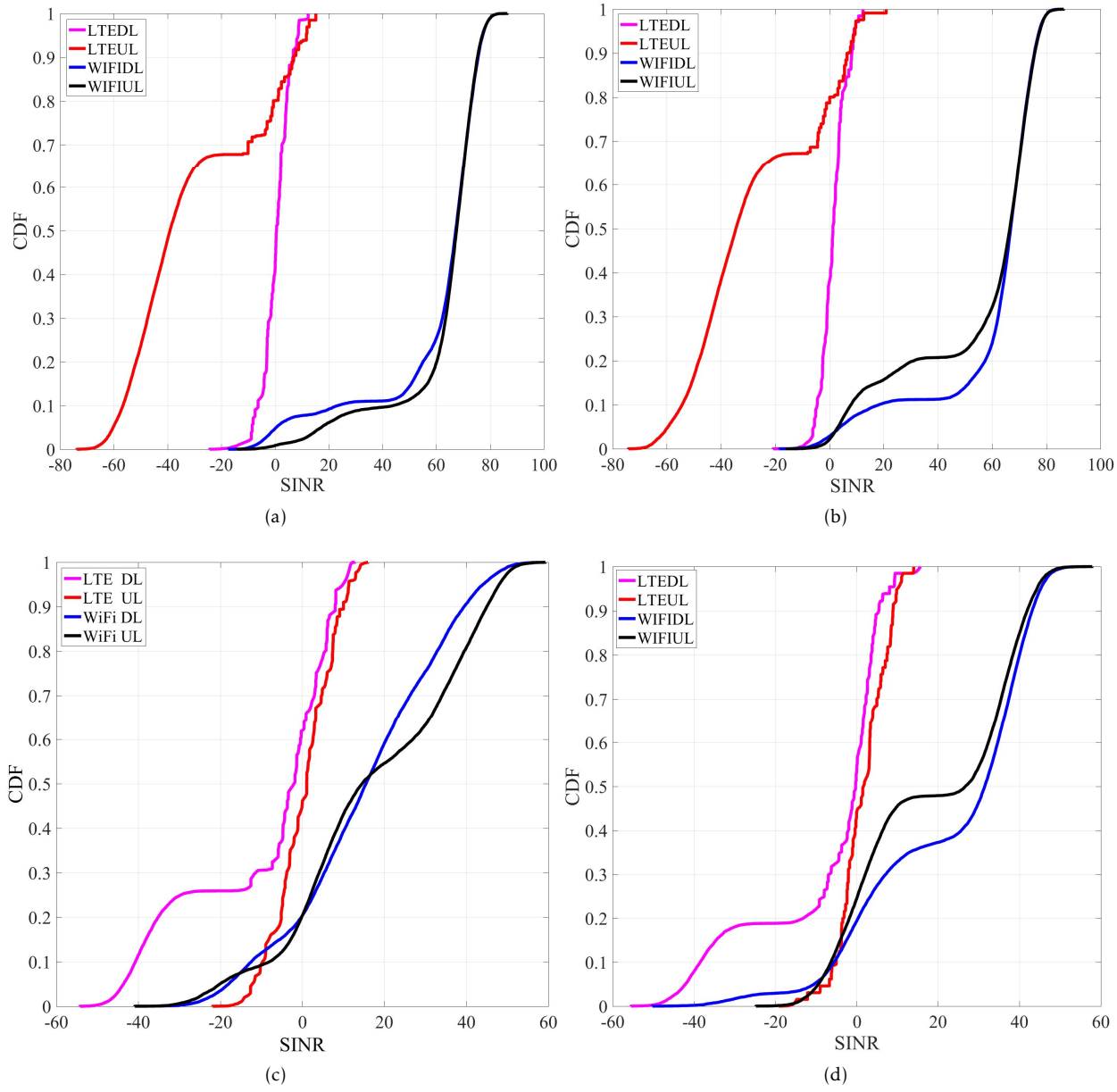


Figure 3: SINR distribution of coexisted LTE-WiFi system (a) SINR distribution at 20% duty cycle (b) SINR distribution at 40% duty cycle (c) SINR distribution at 60% duty cycle (d) SINR distribution at 80% duty cycle

hand, in the rural areas, scattered and limited number of meters will use WiFi. Therefore, in this case, more access can be given to LTE by selecting higher duty cycles such as 60% and 80%.

5 Conclusion

In this study, a collocated WiFi and LTE based advance metering infrastructure is proposed for smart grid. For meter-to-meter data communication, WiFi is proposed. On the other hand, for sending collected data from a group of meters to MDMS, LTE is proposed. A fixed duty cycle of a transmission time is reserved for LTE and the rest of the period is given to WiFi system. The simulation performance shows a harmonious neighborhood spectrum sharing between LTE and WiFi. With the increase of LTE duty cycle, the throughput, energy efficiency and SINR of LTE are improved along with degradation of those of WiFi.

The transmission duty cycle of LTE is adjustable based on the amount of data and number of smart meters. In particular, lower duty cycle of LTE transmission can be selected for urban and suburban areas where the density of smart meters are high and meters need more access to WiFi. On the other hand, higher duty of LTE transmission can be selected for rural areas where the density of smart meter is low. The CBRS band has a big amount of free, underutilized, and clean spectrum for wireless network. So, network consists of coexisting LTE and WiFi in CBRS band can be a viable communication solution for the metering infrastructure of smart grid.

Acknowledgment The work is a outcome of the research supported by the U.S. National Science Foundation under the grant CPS-1446570 and CAREER-1553494.

References

[1] A. F. Molisch, *3GPP LongTerm Evolution*. Wiley-IEEE Press, 2011, pp. 665–698. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5635449

[2] I. Parvez, M. Jamei, A. Sundararajan, and A. I. Sarwat, "RSS based loop-free compass routing protocol for data communication in advanced metering infrastructure (AMI) of Smart Grid," in *2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, Dec 2014, pp. 1–6.

[3] I. Parvez, A. I. Sarwat, J. Pinto, Z. Parvez, and M. A. Khandaker, "A gossip algorithm based clock synchronization scheme for smart grid applications," *CoRR*, vol. abs/1707.08216, 2017. [Online]. Available: http://arxiv.org/abs/1707.08216

[4] M. Jafari, T. O. Olowu, and A. I. Sarwat, "Optimal smart inverters volt-var curve selection with a multi-objective volt-var optimization using evolutionary algorithm approach," in *2018 North American Power Symposium (NAPS)*, Sep. 2018, pp. 1–6.

[5] I. Parvez, A. Islam, and F. Kaleem, "A key management-based two-level encryption method for ami," in *2014 IEEE PES General Meeting — Conference Exposition*, July 2014, pp. 1–5.

[6] I. Parvez, A. I. Sarwat, L. Wei, and A. Sundararajan, "Securing metering infrastructure of smart grid: A machine learning and localization based key management approach," *Energies*, vol. 9, no. 9, 2016.

[7] I. Parvez, F. Abdul, and A. I. Sarwat, "A location based key management system for advanced metering infrastructure of smart grid," in *2016 IEEE Green Technologies Conference (GreenTech)*, April 2016, pp. 62–67.

[8] M. O. Onibonoje and T. O. Olowu, "Real-time remote monitoring and automated control of granary environmental factors using wireless sensor network," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Sep. 2017, pp. 113–118.

[9] T. O. Olowu, M. Jafari, and A. I. Sarwat, "A multi-objective optimization technique for volt-var control with high pv penetration using genetic algorithm," in *2018 North American Power Symposium (NAPS)*, Sep. 2018, pp. 1–6.

[10] I. Parvez, A. Sundararajan, and A. I. Sarwat, "Frequency band for HAN and NAN communication in Smart Grid," in *Proc. IEEE Symp. on Comp. Intel. App. in Smart Grid (CIASG)*, Dec 2014, pp. 1–5.

[11] I. Parvez, A. Rahmati, I. Güvenç, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *CoRR*, vol. abs/1708.02562, 2017. [Online]. Available: http://arxiv.org/abs/1708.02562

[12] 3GPP, "Study on Licensed-Assisted Access using LTE," RP-141397, Tech. Rep., Sep. 2014.

[13] N. Rupasinghe and İ. Güvenç, "Licensed-assisted access for WiFi-LTE coexistence in the unlicensed spectrum," in *Proc. IEEE Glob. Work.(GC Wkshps)*. IEEE, 2014, pp. 894–899.

[14] F. Chaves, A. Cavalcante, E. Almeida, F. Abinader Jr, R. Vieira, S. Choudhury, and K. Doppler, "LTE/Wi-Fi coexistence: Challenges and mechanisms," in *2013, XXXI SIMPOSIO BRASILEIRO DE TELECOMUNICACOES–SBrT2013*, 2013.

[15] FCC, "Amendment of the Commissions Rules with Regard to Commercial Operations in the 3550-3650 MHz Band," Report and Order and Second Further Notice of Proposed Rulemaking, Apr. 2015.

[16] L. Stefani, "The FCC Raises the Curtain on the Citizens Broadband Radio Service," CommLawBlog Article, May 2015. [Online]. Available: http://www.commlawblog.com/2015/05/articles/unlicensed-operations-and-emer/the-fcc-raises-the-curtain-on-the-citizens-broadband-radio-service/

[17] G. Locke and L. E. Strickling, "An assessment of the near-term viability of accommodating wireless broadband systems in the 1675-1710 MHz, 1755-1780 MHz, 3500-3650 MHz, and 4200-4220 MHz, 4380-4400 MHz bands," Report, Oct. 2010. [Online]. Available: http://www.ntia.doc.gov/files/ntia/publications/fasttrackevaluation11152010.pdf

[18] I. Parvez, M. Sriyananda, İ. Güvenç, M. Bennis, and A. Sarwat, "Cbrs spectrum sharing between lte-u and wifi: A multiarmed bandit approach," *Mobile Information Systems*, vol. 2016, 2016.

[19] S. Rahman, M. Moghaddami, A. I. Sarwat, T. Olowu, and M. Jafaritalarposhti, "Flicker estimation associated with pv integrated distribution network," in *SoutheastCon 2018*, April 2018, pp. 1–6.

[20] I. Parvez, N. Chotikorn, and A. I. Sarwat, "Average quantized consensus building by gossip algorithm using 16 bit quantization and efficient data transfer method."

[21] Y. Mekonnen, M. Haque, I. Parvez, A. H. Moghadasi, and A. I. Sarwat, "LTE and wi-fi coexistence in unlicensed spectrum with application to smart grid: A review," *CoRR*, vol. abs/1708.09005, 2017. [Online]. Available: http://arxiv.org/abs/1708.09005

[22] I. Parvez, T. Khan, A. I. Sarwat, and Z. Parvez, "Laa-lte and wifi based smart grid metering infrastructure in 3.5 ghz band," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec 2017, pp. 151–155.

[23] X. Wang, S. Mao, and M. X. Gong, "A survey of lte wi-fi coexistence in unlicensed bands," *GetMobile: Mobile Computing and Communications*, vol. 20, no. 3, pp. 17–23, 2017.

[24] A. Babaei, J. Andreoli-Fang, Y. Pang, and B. Hamzeh, "On the impact of LTE-U on Wi-Fi performance," *International Journal of Wireless Information Networks*, vol. 22, no. 4, pp. 336–344, 2015.

[25] M. Sriyananda, I. Parvez, I. Güvene, M. Bennis, and A. I. Sarwat, "Multi-armed bandit for LTE-U and WiFi coexistence," in *Proc. IEEE Wire. Commun. and Netw. Conf.* IEEE, 2016, pp. 1–6.

[26] G. RP-140808, "Review of regulatory requirements for unlicensed spectrum)," AlcatelLucent, Alcatel-Lucent Shanghai Bell, Ericsson, Huawei, HiSilicon, IAESI, LG, Nokia, NSN, Qualcomm, NTT Docomo, Tech. Rep., June 2014.

[27] Nokia, "Nokia LTE for unlicensed spectrum," AlcatelLucent, Alcatel-Lucent Shanghai Bell, Ericsson, Huawei, HiSilicon, IAESI, LG, Nokia, NSN, Qualcomm, NTT Docomo, Tech. Rep., June 2014.

[28] F. M. Abinader, E. P. L. Almeida, F. S. Chaves, A. M. Cavalcante, R. D. Vieira, R. C. D. Paiva, A. M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler, and V. A. Sousa, "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54–61, Nov. 2014.

[29] Qualcomm, "Qualcomm Research LTE in Unlicensed Spectrum: Harmonious Coexistence with WiFi," 3GPP RAN1 standard contribution, Tech. Rep., Jun. 2014.

[30] ZTE, "Frame structure design for LAA considering LBT," 3GPP TSG RAN WG1 Meeting no.80, Feb. 2015.

[31] R. Zhang, M. Wang, L. X. Cai, X. Shen, L.-L. Xie, and Y. Cheng, "Modeling and analysis of MAC protocol for LTE-U co-existing with Wi-Fi," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*. IEEE, 2015, pp. 1–6.

[32] O. El-Samadisy, M. Khedr, and A. El-Helw, "Performance Evaluation of MAC for IEEE 802.11 and LAA LTE," in *Proc. Intern. Conf. on Comput. Science and Comput. Intell. (CSCI)*, Dec 2016, pp. 923–928.

[33] A. k. Ajami and H. Artail, "Fairness in future licensed assisted access (laa) lte networks: What happens when operators have different channel access priorities?" in *Proc. IEEE Intern. Conf. on Commun. Works. (ICC Workshops)*, May 2017, pp. 67–72.

[34] R. Ratasuk, M. A. Uusitalo, N. Mangalvedhe, A. Sorri, S. Iraji, C. Wijting, and A. Ghosh, "License-exempt LTE deployment in heterogeneous network," in *Int. Symp. on Wireless Commun. Sys.*, Paris, France, Aug. 2012, pp. 246–250.

[35] E. Almeida, A. M. Cavalcante, R. C. D. Paiva, F. S. Chaves, F. M. Abinader, R. D. Vieira, S. Choudhury, E. Tuomaala, and K. Doppler, "Enabling LTE/WiFi coexistence by LTE blank subframe allocation," in *Proc.IEEE Int. Conf.*, Budapest, Hungary, Jun. 2013, pp. 5083–5088.

[36] T. Nihtil, V. Tykhomyrov, O. Alanen, M. A. Uusitalo, A. Sorri, M. Moisio, S. Iraji, R. Ratasuk, and N. Mangalvedhe, "System performance of LTE and IEEE 802.11 coexisting on a shared frequency band," in *Proc. IEEE Wireless Commun. and Net. Conf.(WCNC)*, Shanghai, China, Apr. 2013, pp. 1038–1043.

[37] I. Parvez, N. Islam, N. Rupasinghe, A. I. Sarwat, and . Gven, "LAA-based LTE and ZigBee coexistence for unlicensed-band smart grid communications," in *Proc. IEEE SoutheastCon conf.*, March 2016, pp. 1–6.

[38] F. Granelli, D. Domeniconi, N. L. S. D. Fonseca, and B. Tsetsgee, "On the usage of wifi and lte for the smart grid," in *2014 7th International Conference on Ubi-Media Computing and Workshops*, July 2014, pp. 1–5.

[39] F. Koohifar, N. Saputro, I. Guvenc, and K. Akkaya, "Hybrid wi-fi/lte aggregation architecture for smart meter communications," in *2015 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Nov 2015, pp. 575–580.

[40] N. Rupasinghe and . Gven, "Licensed-assisted access for wifi-lte coexistence in the unlicensed spectrum," in *2014 IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 894–899.

[41] I. Parvez, N. Islam, N. Rupasinghe, A. I. Sarwat, and . Gven, "LAA-based LTE and ZigBee coexistence for unlicensed-band smart grid communications," in *SoutheastCon 2016*, March 2016, pp. 1–6.

[42] M. G. S. Sriyananda, I. Parvez, I. Gvene, M. Bennis, and A. I. Sarwat, "Multi-armed bandit for LTE-U and WiFi coexistence in unlicensed bands," in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[43] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)," 3GPP Technical Report, Tech. Rep., Mar. 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 73-83 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

State Estimation based Echolocation Bionics andImage Processing based Target Pattern Recognition

David Kondru[*,1], Mehmet Celenk[1], Xiaoping A. Shen[2]

[1]*School of Electrical Engineering and Computer Science,Ohio University, Athens, Ohio-45701, USA*

[2]*Department of Mathematics, Ohio University, Athens, Ohio-45701, USA*

ARTICLE INFO

Article history:
Received: 28November, 2018
Accepted: 08 January, 2019
Online : 20 January, 2019

Keywords:
Bionics
Kalman filter
Detection

ABSTRACT

This paper deals with the theoretical aspect of bat echolocation and bionics, and image processing-based target recognition and identification methods. The state estimation methods utilizing the linear rustic filters such as fixed gain and Kalman filters are studied and implemented for echolocation bionics for estimating the LOS distance. A complete mathematical modeling and simulation of bat dynamics and its prey are presented upon which the relative LOS distance is reconstructed with state estimators and less RMS errors. Also, target recognition and identification using Optical, IR Digital Night Vision and Thermal camera is studied and implemented at different environmental conditions to demonstrate the superiority of thermal camera.

1. Introduction

Since the dawn of the Strategic Defense Initiative (SDI), different types of threats and the complex defense system architectures have always been a confrontation to the target detection and tracking. In this context, detection is always prior to tracking without which the aerospace guidance and control is practically impossible [1]. In fact, this is true for civilian sensors such as Primary and Secondary Surveillance Radars to control incoming and departing airplanes, as well as, for military sensors to perform a sequence of successful tasks [2]. The degree of target detection and tracking varies from single sensor to multiple sensors which led to multi sensor fusion for multi object tracking. Intelligence, Surveillance and Reconnaissance (ISR), missile guidance and control, remote sensing and oceanography, computer vision and robotic applications are the few wide variety of applications utilizing from single sensor to multiple sensor detection and tracking [3]. Based on the type of threat and nature of attack many integrated solutions have been introduced to confront with the threat. One such an example is U.S. National Missile Defense system. Aurgus and Cerebrus are the FLIR designed integrated systems for border surveillance with 24/7 situational awareness [4]. Apart from RADAR, Optical and Thermal technology, Bat acoustic detection system, a new technology was introduced for target localization and identification. Echolocation bionics [5], [6] based on the bat technology utilizing the acoustic echo is a process of determining the relative range and velocity, bearing angle and size of the target. In fact, echolocation bionics has the superior advantages and success rate compared to other detection and tracking technologies that made it helpful in civil and military applications. However, irrespective of any given dynamic situation and the type of technology, the designated track has to be followed by a closed loop system tracker for airborne interception or ground attack that requires high performance data association algorithms. The purpose of data association and processing algorithms is to mitigate the high degree of uncertainty associated with the target motion and the environmental conditions resulting from system noise and measurement noise [7] [8]. With these insights, this paper presents the theoretical advancements in echolocation bionics and, tracking and filtering of bat acoustic signal that could enhance the performance of echolocation bionic sensors. On the other hand, this paper also focuses on application of image processing techniques using Finite Impulse Response filters as an added extension to the echolocation bionics. Two-dimensional FIR filters are the most profoundly used filters for image data processing from a given particular optical device. Image data processing techniques such as image sharpening and smoothing, detecting edges in a particular image clustered with many features, and contrast improvement are necessary to apply for agiven 2D image data in the presence of noisy environments. All these techniques are applied to visualize the data as being the key objective in modern computational sciences. Data visualization is not an easy operation as how it has been in a perfect daylight conditions. In poor visibility, capturing and processing of image data and its

[*]David Kondru, Ohio University, Email: k.rajusolomon@gmail.com

visualization is quite daunting and a difficult task. However, there are optical and thermal sensors that can enhance processing of an image even in poor visibility conditions. In this paper, the image of a drone is captured by a digital night vision camera with built-in IR mode and a thermal camera to examine the effects with a 2D FIR filter. This paper presents the superiority and the performance of individual sensor in detection and enhancement, and recognize the patterns in the captured images. Section 2 presents the echolocation bionics and state estimation. Section 3discusses the sensor specifications and test images of target. Section 4demonstrates the image processing based target recognition and identification methods.

2. Echolocation and Tracking

Usually, the mouth (nose) of bat is for broadcasting echolocation and its ears were used as receiving antennas. The emission system of bat is an adaptive waveform where each pulse consists of eight signals. Four long-constant frequency (CF) and four short constant frequency modulated (FM) harmonics [9]. The speed information is obtained by CF component and the FM component determines the detection and imaging. Whenever a bat gets closer to its stationary prey, due to Doppler Effect the echo frequency becomes higher than the emitted pulse [10]. At the same time, the pulse density becomes larger and larger which is shown in Figure 1 describing the attack process of a bat and the adaptive waveform conversion on the top left.

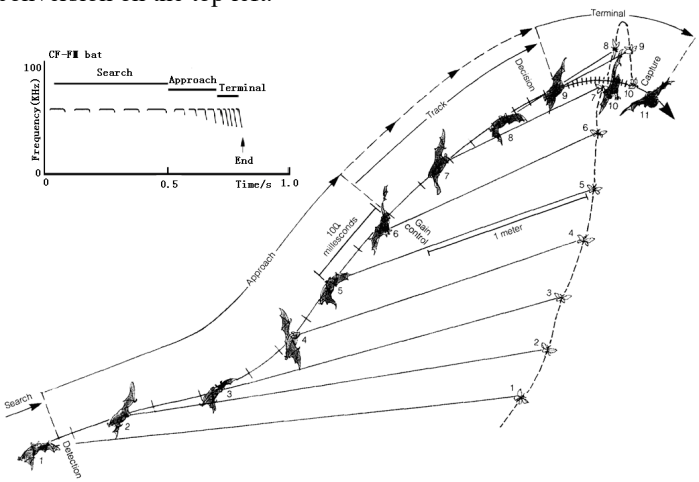


Figure 1.Bat intercepting its prey [10]

A three dimensional simulation of a bat sonar beam while attacking its prey, and the two dimensional spectrogram sequence of the bat echolocation relative to the target [11] is shown in Figure 2. Initially the bat searches for its prey using its directional beam and once the prey is detected the target is locked with the sonar beam. The bat while hunting to its prey decreases the number of echoes it emits using automatic gain control.

For the purpose of simulation and based on the literature available as represented in the Figures 1 and 2, the three dimensional motion of the bat is considered as a spiral motion and the prey or the target is stationary. The position and the velocity of the target are almost negligible when compared to the motion of the bat. It is assumed that the bat has already received the passive observation of the target and the simulation shows the hunting of the bat toward its prey. The relative position which is also known as line of sight (LOS) between the bat and the prey is considered for tracking using different state estimators and the performance evaluation of each estimator in terms of RMS error measure. Using the flat earth approximation [12], the target and its prey are described in a Spherical Coordinate system [13] upon which the position is illustrated in Cartesian coordinate system for 3D target tracking. Therefore, the finite dimensional representation [14] for the target and the prey model, as well as, the LOS system measurements [15] are modeled by

$$[x\ \dot{x}\ddot{x}\ y\ \dot{y}\ \ddot{y}\ z\ \dot{z}\ \ddot{z}]^T \tag{1}$$

$$r = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{2}$$

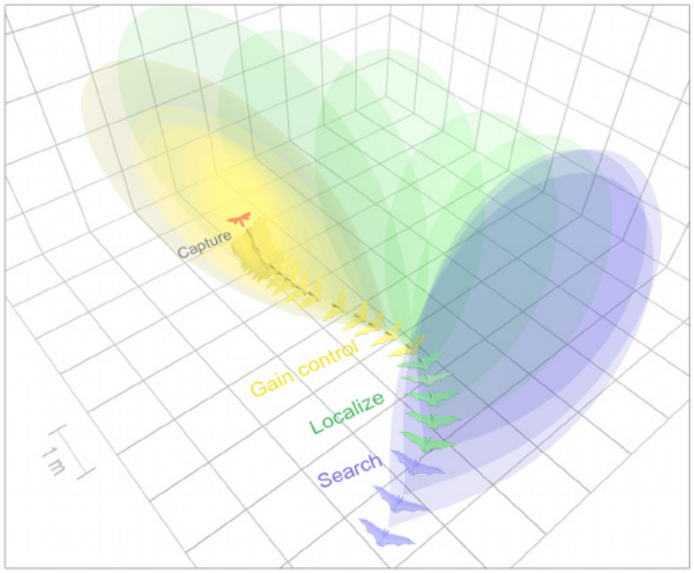


Figure 2. Approach methods of bat [11]

Equation 1 describes the 3D motion characteristics of bat and its prey in terms of position, velocity and acceleration. Equation 2 presents the relative position or the line of sight distance between the bat and its target. Figure 3 shows the motion characteristics [16] that are utilized in state variable form and all the motions are simulated in MATLAB.

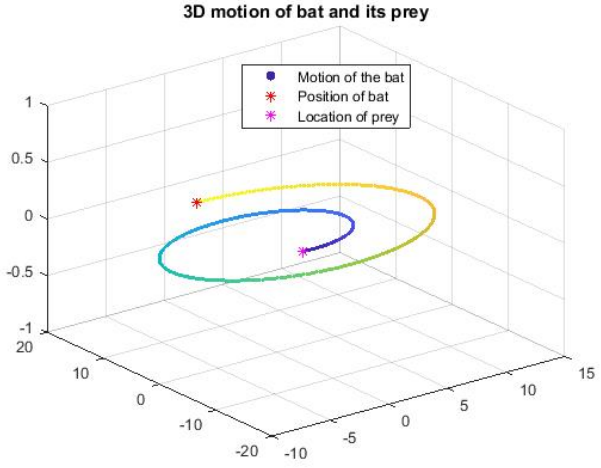


Figure 3. Bat its prey motion dynamics

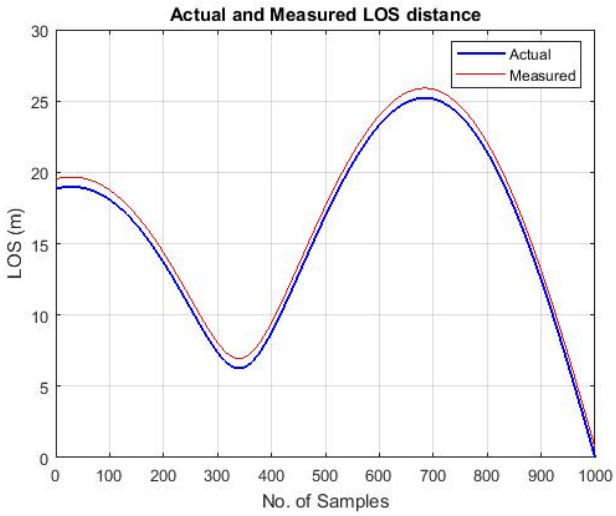


Figure 4.Relative LOS distance

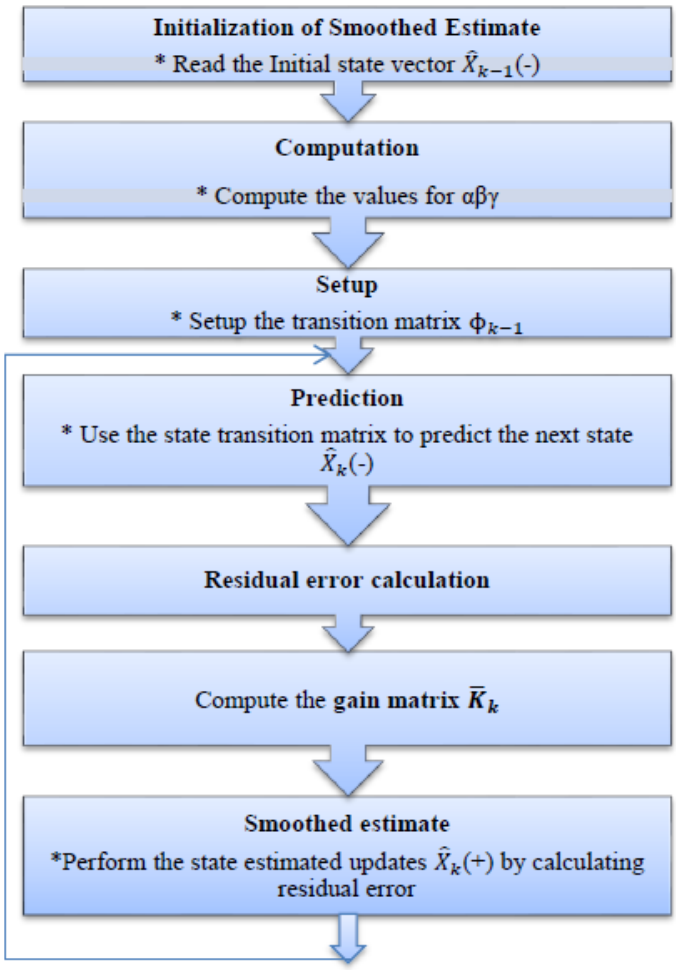


Figure 5. Fixed gain filter recursive flow diagram [19]

2.1. State Estimation of Line of Sight

The difference in position between the bat and its prey is considered as the line of sight (LOS) distance for which a popular finite-dimension approach known as state estimation is utilized in this paper for predicting and correcting the LOS distance. Therefore, the generic representation of a kinetic model [17] is given by

$$\hat{X}_k(-) = F\hat{X}_{k-1}(+) + Gw_{k-1} \quad (3)$$

$$Z_k = H\hat{X}_{k-1}(+) + v_k \quad (4)$$

Two different filters namely fixed-gain andKalman filters are studied and implemented for LOS estimation. Before examining each filter, the underlying assumption made is portraying the measurement acoustic noise as zero-mean Gaussian (ZMG) as shown in the Figure 4 with a known standard deviation.

2.2. Linear Rustic Filters

Assuming that the system is a linear time invariant [18], the first class of rustic filters namely fixed gain and Kalman filters are implemented for which the process flow diagram for each filter is shown in Figures 5 and 6. A one dimensional third order filter is adopted for the LOS estimation.

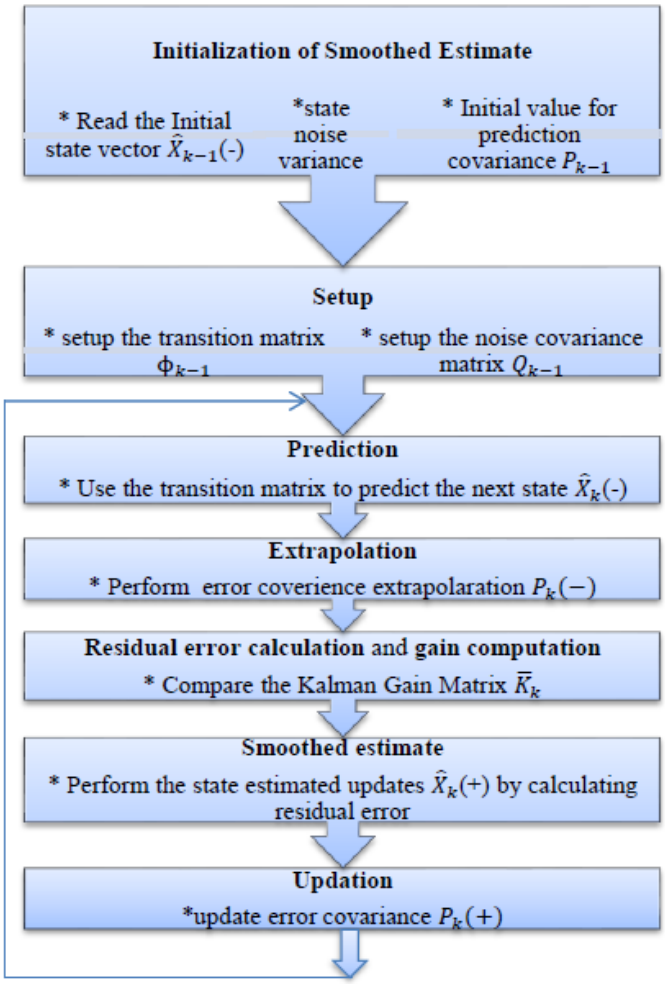


Figure 6. Kalman filter recursive flow diagram [19]

Table 1 Simulation Parameters

Sampling Interval (T)	0.01 sec
Measurement Noise Variance w_K	0.035
Process Noise Variance v_k	0.5

2.3. Simulation Analysis

For the purpose of simulation, it is assumed that both bat and prey are in the vicinity with a radius of maximum of 25 meters. The simulation parameters are shown in Table1.

The initial values of prediction covariance and transition matrix are employed as follows:

$$P_{k-1}=\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}; \qquad \phi = \begin{bmatrix}1 & T & T^2/2\\ 0 & 1 & T\\ 0 & 0 & 1\end{bmatrix} \tag{5}$$

To ensure the performance of each estimator the error analysis in LOS distance is computed using RMS error given by the Equation 6.

$$RMS\ Error_{pos,ang} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(x_{true}-x_{esti})^2+(y_{true}-y_{esti})^2+(z_{true}-z_{esti})^2}{3}} \tag{6}$$

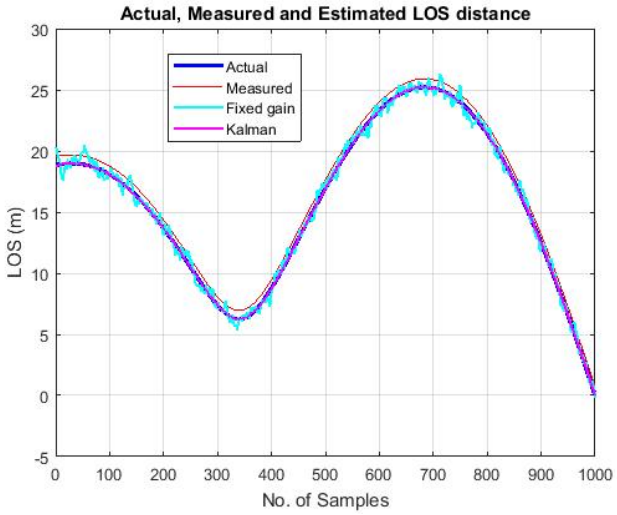


Figure 7. Actual, Measured and LOS distances

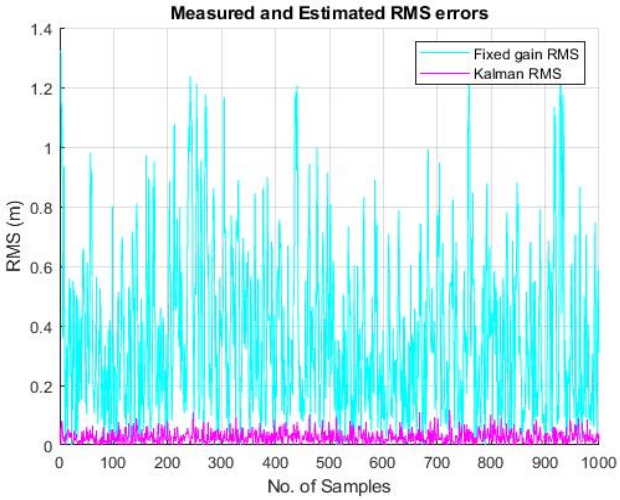


Figure 8. RMS Errors

Figure 7 shows the tracking performances for LOS distance. The principal function of these trackers is for better signal reconstruction by reducing measurement noise with less residual. The performance of each filter estimates is shown from Figure 8. It is found that the efficiency of fixed gain filter is highly reliable upon the filter gain coefficients which are solely depend upon the smoothing coefficient that varies from zero to one. A value of 1 is chosen for heavy smoothing and low value for less smoothing. Whereas the Kalman filter has a better estimate than fixed gain filter because the Kalman gains are computed dynamically. In other words the Kalman gain matrix is determined from the state error variance as well as from measurement and noise variances adjusts adaptively. The Kalman filter computes its own state error uncertainty estimates while observing a new measurement. Also as it is mentioned [20] if the new measurement noise error variance is bigger than the state error variance, the Kalman filter will place less emphasis on the new measurement than if the state error is larger than the measurement error. On the other hand the simple fixed gain filter computes the filter gains based on assigned smoothing coefficient. This effect and the superiority of Kalman filter over fixed gain estimator can be seen in RMS errors shown in Figure 8. The performance of the fixed gain and Kalman filters serves as best linear estimators for reducing mean squared errors in LOS distance. The next section deals with the target recognition and identification using image processing techniques. From the Figure 8 and using the equation 6 the RMS errors for measured, fixed gain and Kalman filter is found to be 1.5350 m, 0.3119 m and 0.0206 m, respectively, and are shown in Table 2.

Table 2: Thermal Camera Specs

LOS RMS errors	
Filter type	**RMS error**
Measured	1.5350
Fixed gain filter	0.3119
Kalman filter	0.0206

Table 3: IR Digital Night Vision Camera Specs

Sensor Type	Technical Specs	
Digital Night Vision	Sensor	CMOS Sensor, Max. 24 MP
	Lens	Fixed lens, F/3.2, f=7.5mm
	Focus Range	Normal: 1m ~ infinity
	Image	Resolution: FHD, 3Mp – 24 Mp
	Video	Resolution: FHD (1920×1080) - VGA
	Night Vision range	<3.0 m
IR LED Illuminator	No. of LED	198
	IR range	300-400 feet (100-130 meters)
	Light beam angle	45-60 degree
	Wavelength	850nm

3. Sensor Specifications

3.1. Night Vision Camera

A night vision digital video camera is employed for capturing the images to post process the images in the MATLAB. The night vision camera selected for this experiment belongs to a class of image intensifier (I^2) device. Also, a CMVision IR infrared

illuminator is added to the night vision camera for more illumination of the target drone. The important technical specifications [21] of the two devices are given in Table 3.

3.2. Thermal Camera

A Seek CompactXR thermal imaging camera designed for smart phone is utilized in this research work. The purpose of using a thermal camera is based on their ability to detect and track in extreme pitch black conditions.They are more sensitive to temperature variations. Designed with over 32,000 thermal pixels, the CompactXR will sense the temperature illumination and displays on the smart phone installed with the Seek thermal app. The important technical features [22] of the thermal sensor are given in Table 4.

Table 4: Thermal Camera Specs

Technical Specs	
Sensor	206×156 thermal sensor
Field of View (FoV)	20^0 narrow FoV
Operating distance	1,800 ft (548 m)
Detection range	-40^0F to 626^0F
Wavelength	Long Wave Infrared (7.2 - 13 Microns)

3.3. Target Drone and Platform Setup

A 4 channel X5C 2.4GHz remote control quadcopter is used as a target drone. The size of the drone is 16.5 x 12.2 x 3.8 inches. Although, this drone is designed with built-in 6 axis gyroscope with 360^0 3D eversion and throwing flight function, the target drone was made to hang on a wallas well as tree for the sake of simplicity in simulation. A tripod mounted with different sensors shown in Figure 9 is used to capture the images of the drone at different light conditions.

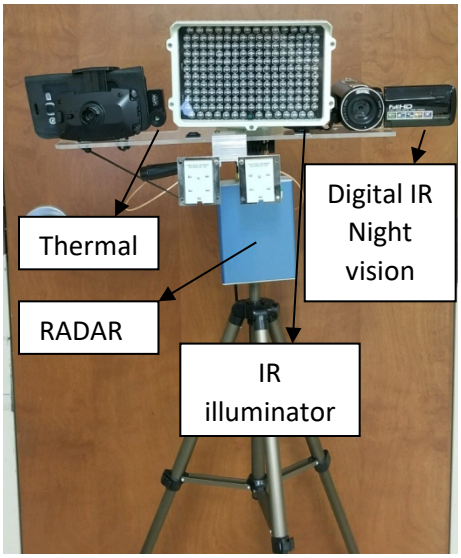


Figure 9. Sensors Tripod

To better understand the image processing techniques, two different environmental backgrounds are considered and the performance of each device is observed. The first set of images is taken with a dark background and the second set of images is taken as the target drone blended to the outside environment.

A. First set of Images

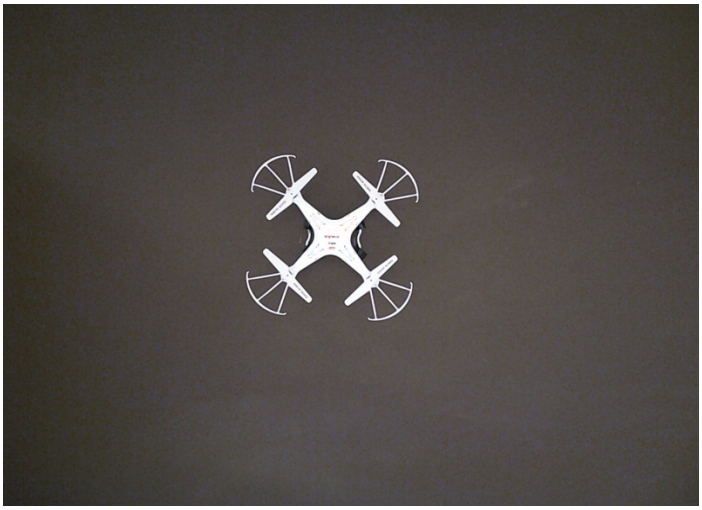


Figure 10. Test image with dark background

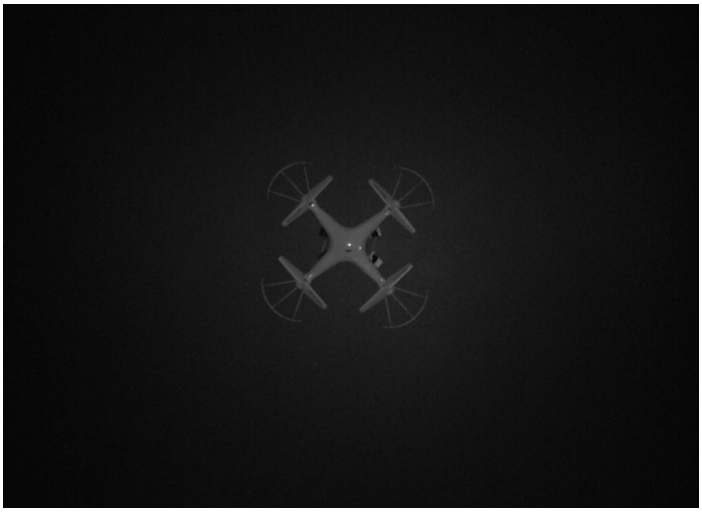


Figure 11. Test image in night vision

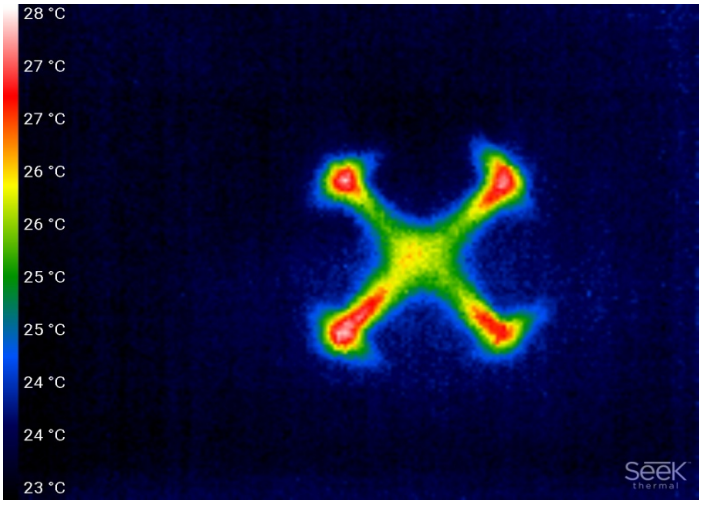


Figure 12. Thermal image

B. Second Set of Images

Figure 13. Test image when target blended to the background

Figure 14. Figure: Test image in night vision

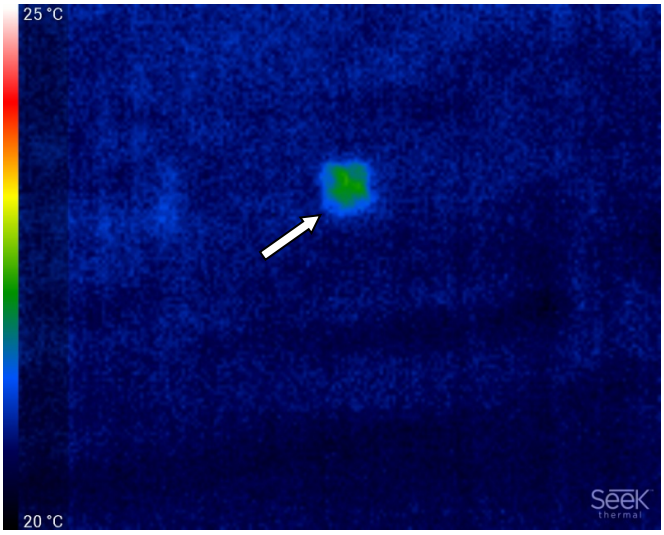


Figure 15. Thermal image

The heat dissipated by the motors of the drone can be easily seen in Figures 12 and 13. The next section deals with the numerical simulation analysis by utilizing the edge detection, image sharpening, smoothing, and contrast enhancement techniques [20].

4. Numerical Simulation Analysis

Each pixel element representing the brightness on the corresponding location in an image is an 8-bit number and can be combined with the adjacent pixels through simple weighted average. Also, the texture of the image as well as sharpness, resolution and contrast can be altered. Moreover, the importance of feature extraction using a signal processing filter known as template detector can also be applied to an image to extract the features like edges of a drone when blended into the surrounding clutter. For the regions where there is a perfect match between the template and pixel, the template detector produces high output for those regions. In the neighborhood of diversity of drones, the computer will be able to recognize one drone from the other using vision feature in which the 2D digital filter or the convolver has become the main processing tool. The core function of a 2D filter is to combine the local pixels with the weights to output the filtered image brightness. It can be expressed as

$$B'(x,y) = \sum_{i=-N}^{+N} \sum_{j=-M}^{+M} w_{i,j} B(x+i, y+j) \quad (7)$$

where, $B'(x,y)$ is the processed image brightness and $w_{i,j}$ are the weights. Based on the solid foundation of 2D FIR filters, the moving average technique has originated as a fundamental process for brightness computation that utilizes weighted average of the neighborhood pixels. The visual effects such as alteration of image focus, image contrast and sharpness can also be implemented with 2D filters. They can enhance and reduce the texture features in an image. One of the simple illustrations of a 2D FIR filtering process is pixilation or pixel decimation where all the pixels are replaced by the average of the pixels of original higher resolution image. This technique is often applied in television broadcast to obscure some offensive images and blur the selected areas.

4.1. Edge Detection

One of the wider applications of a 2D FIR filter is edge detection to enhance boundaries or the shape of an object in an image.As 2D FIR filtering being a prominent application in computer pattern recognition, determining the edges of a target in a particular image can extract important features such as its size and shape. The target considered here is a drone in the image captured by digital night vision and thermal camera. A gradient filter which is an example of an edge detector is applied in the direction of positive x-gradient. Since it is undesirable to produce an output image offset by half a pixel in the desired gradient direction, a symmetric finite-difference approximation is typically used to avoid the unwanted offsets. It is given by the equation

$$B'(x,y) = -\frac{\partial}{\partial x} B(x,y)$$

$$= -[B(x+1,y) - B(x-1,y)] \quad (8)$$

The effect of the negative sign in front of the equation 8 is shown in Figures 16, 18 and 19. These are the observations performed by

an edge detector using a gradient with directional derivatives. The disadvantage of these directional derivatives is the presence of the inherent noise. As this method is solely based on a finite difference method, there will always be a random noise associated because ofthe difference computation between the two pixels. The finite difference method tends to amplify the noise and averaging the pixels tends to smooth out the image noise. However, apart from these deficiencies, the noise can be suppressed by adding more pixels in the filter. This can be accomplished by simply averaging the derivatives based on the choice of direction. If it is in x-direction, the derivatives in the adjacent rows will be averaged, else the adjacent columns in y-direction. The choice of filter weights when the gradient is in the negative x-direction are given by the equation

$$-\nabla_x = \begin{bmatrix} w_{-1,+1} & w_{0,+1} & w_{-1,+1} \\ w_{-1,0} & w_{0,0} & w_{+1,0} \\ w_{-1,-1} & w_{0,-1} & w_{+1,-1} \end{bmatrix} = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \quad (9)$$

Equation is known as kernel operator that has an ease of rotation based on the desired gradient direction. In essence, the negative kernel operator in upward (positive) y-direction is given by equation

$$-\nabla_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \quad (10)$$

There are different 2D based gradient kernel operators available with respect to size and complexity, and finite-difference approximation based higher order filters. Any two-dimensional image data with gradient information is useful for automatic detection to identify and extract the geometric features. These directional derivative gradients applied either in the x (toward right) and y (upward) axis are essential for detecting the edges in a particular orientation of an image. However, the complete geometric features of an image in all directions, i.e., all edges of a particular object in an image are needed for the application of this paper. Indeed, by detecting all edges simultaneously, the important features such as shape and size, as well as orientation can be determined. A widely used conventional edge detection technique known as Sobel edge detection is utilized in this paper for identifying the target object which is the drone in the test images. A Sobel edge detector first calculates the spatial derivatives in each x-and-y directions. Thereby, it sums their squares and computes the square root. This complex process can be avoided by Kirsh operator which is very less complex in nature. Instead of computing the square root of the sum, the Kirsh operator estimates the neighborhood gradients and calculates the maximum absolute value to produce an edge detected output. Figures 16, 17 and 18 show the application of Sobel edge detection combined with Kirsh operator approximation depicting the significance of edge detection for feature extraction. Moreover, apart from the various features of these images that are useful for pattern recognition analysis, these edge detection techniques are also applicable to enhance the image visual quality.

4.2. Image Enhancement

Typically, these edge detectors are considered as high-pass filters through which the random changes in spatial brightness could be seen. On the other aspect, the sharpness of an image to its highest acuity can be enhanced by amplifying the higher frequencies and attenuating the lower frequencies. A rotational invariant Laplacian operator is applied for the test images shown in Figure 10, 11, 13 and 14 to perform this operation and is given by

$$-\nabla^2 B = \frac{\partial^2 B}{\partial x^2} + \frac{\partial^2 B}{\partial y^2} \quad (11)$$

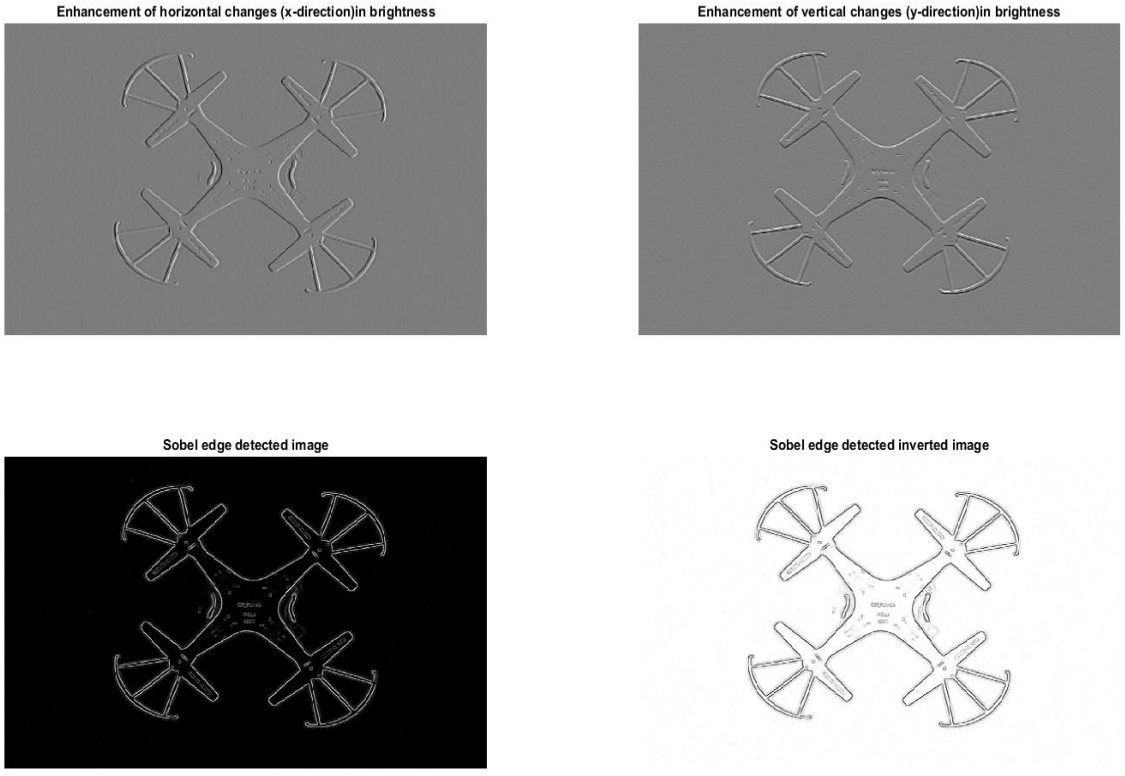


Figure 16. Enhancement and edge detection of optical image

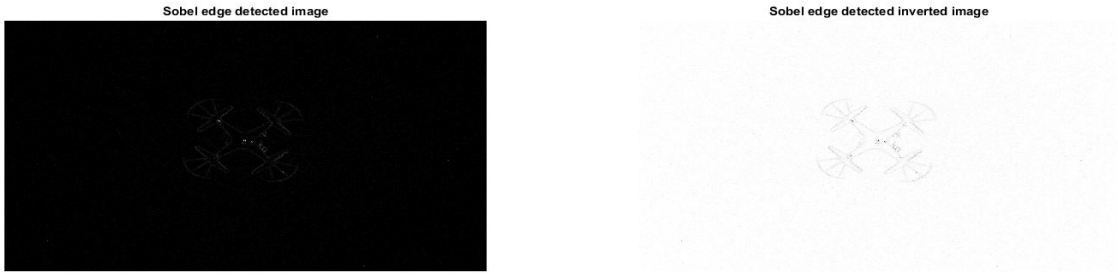


Figure 17. Edge detection of night vision image

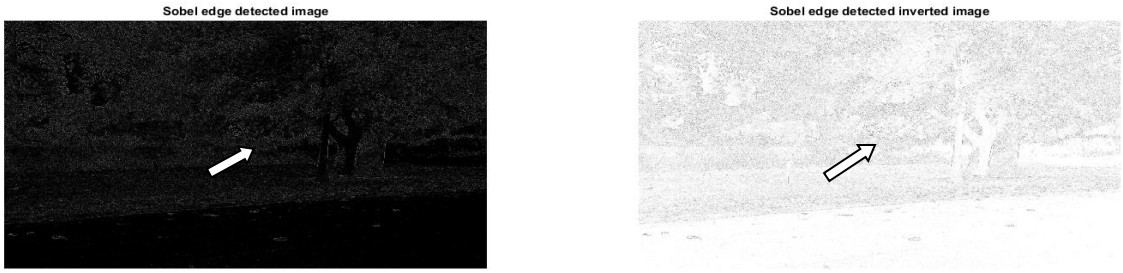


Figure 18. Edge detection of optical image

Figure 19. Edge detection of night vision image

The Laplacian operator employs finite difference method in summing up the x-direction and negative (second) derivatives in y-direction expressed as

$$-\nabla^2 B = \begin{bmatrix} 0 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad (12)$$

In order to reduce noise by considering all eight neighboring pixels as is the case in the edge detector, the addition in the diagonal components are done to negative Laplacian given by

$$-\nabla^2 B = \begin{bmatrix} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (13)$$

Based on the finite differences, equations 11, 12 and 13 are known as normalized operators for approximating a negative Laplacian operator. The normalization was done in such a way that the average brightness in the image is not shifted. The advantage with the Laplacian operator is that it will cancel the varying brightness however, it makes the image noisier as the random variations are amplified by the operator from pixel to pixel. Moreover, the edge enhancement and the image sharpening are done by adding the output image of the negative Laplacian operator to its original image. Although, it adds noise, the high frequency spatial components are enhanced by image sharpening filter, though the filtered image doesn't look like the original image. Hence, in this case an unsharp filter can be utilized to improve the sharpness of an original image by adding or subtracting some percentage to the original image.

4.3. Image Sharpening

The digital un-sharp filter is done either by adding or subtracting the Laplacian filtered image to the original or by subtracting the low-pass-filtered image from the original. The image which is added or subtracted to the original in place of the Laplacian image is considered as the high frequency image. Another less noisy approach is the Laplacian of Gaussian (LOG) filter in which first the image is low pass filtered to remove noises and Laplacian is applied. In essence, as the 2D convolution operation of the filter is associative, the LOG smoothing filter result can be used as the 2D FIR filter. The LOG function is given by

$$\tilde{N}^2 G = \frac{-1}{\pi\sigma^4}\left(1 - \frac{x^2+y^2}{2\sigma^2}\right) e^{\frac{-(x^2+y^2)}{(2\sigma^2)}} \quad (14)$$

where, x and y can vary from -1 to +1 pixel. Assuming σ=0.66, the unshrpen operator is given by

$$\nabla^2 G = \begin{bmatrix} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (15)$$

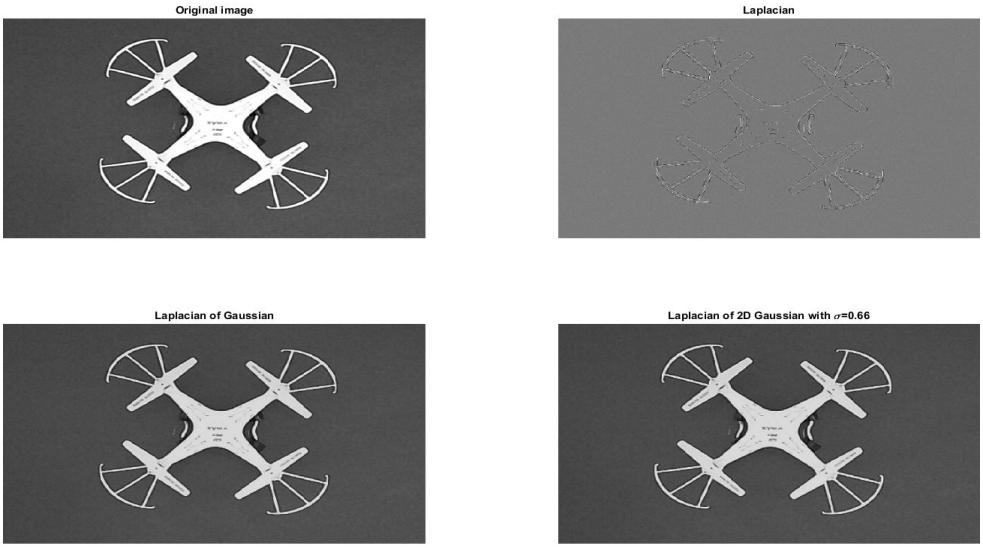


Figure 20. Image enhancement of optical image

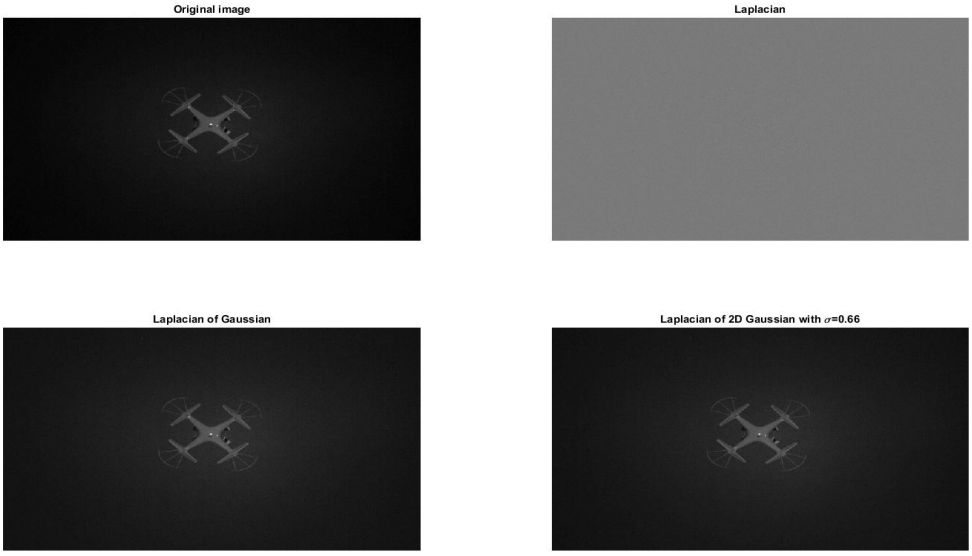


Figure 21. Image enhancement of night vision image

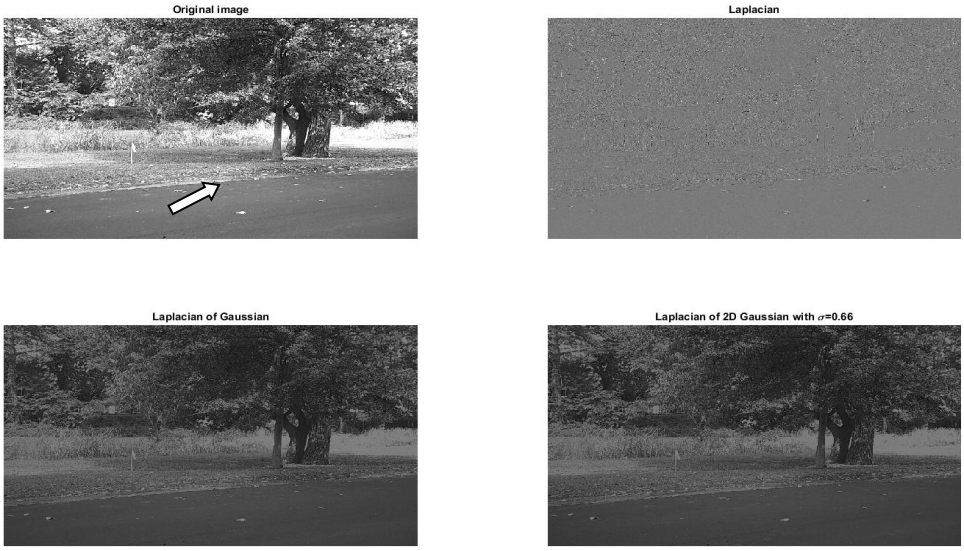


Figure 22. Image enhancement of optical image

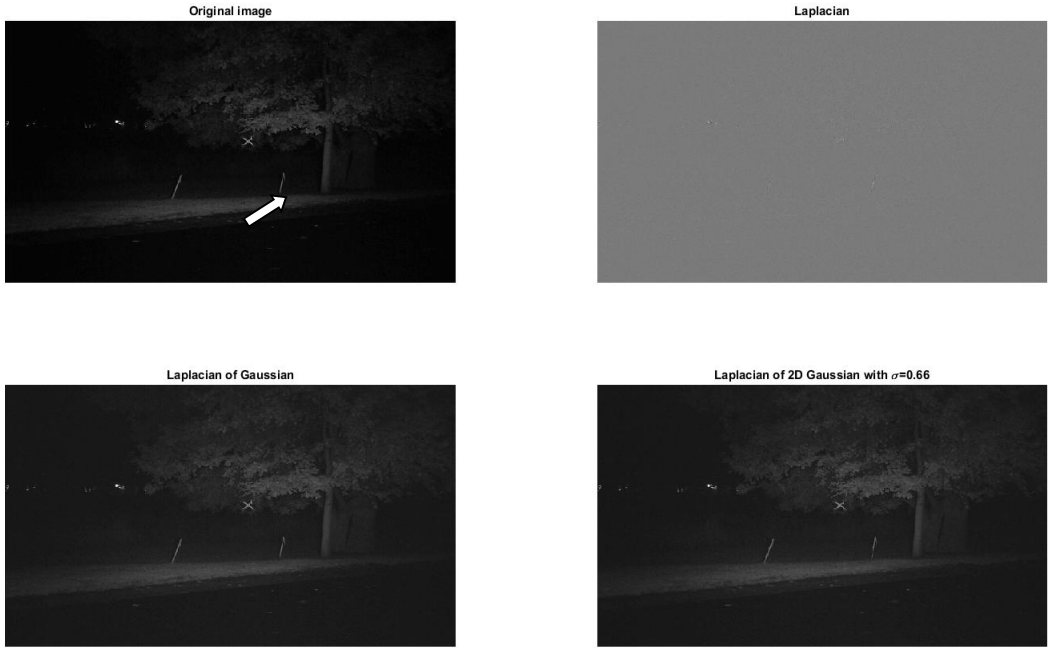


Figure 23. Image enhancement of night vision image

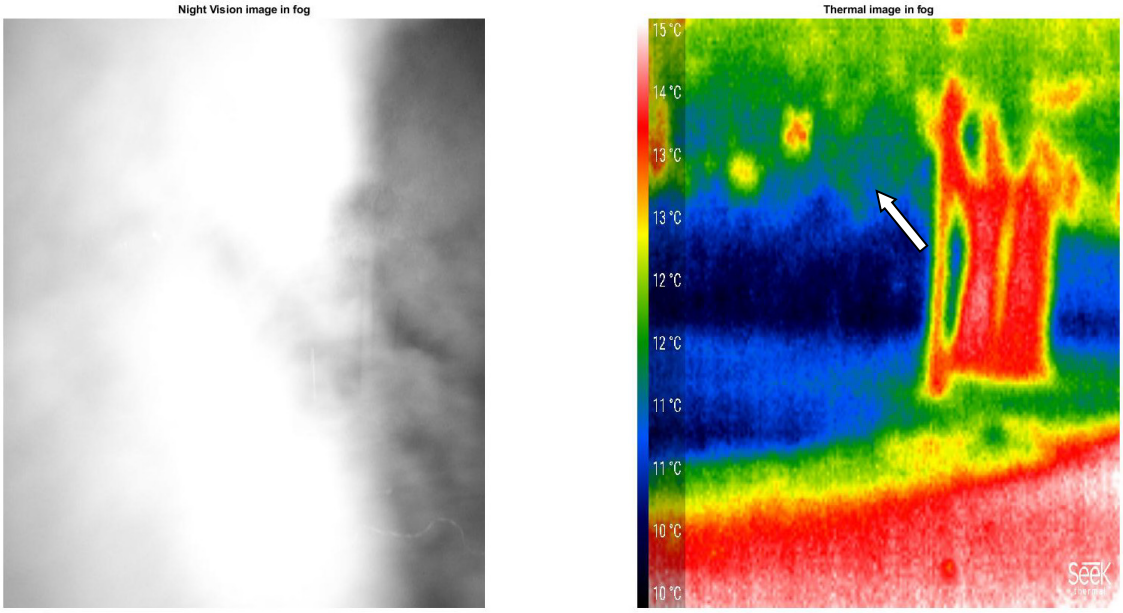


Figure 24. Comparison of night vision and thermal vision in dense fog

Conclusion

This paper deals with the advancement of echolocation bionics and its importance for analyzing and preying upon the target. The mathematical modeling and simulation of bat and its prey is developed upon which the tracking of LOS distance using different state estimators is presented. The RMS error shows the superiority of state estimators for better reconstruction of the signal with less residual. Also, the second half of the paper presents the image processing based techniques to identify the target drone in different places at different environmental situations using multiple sensors. Figure 24 demonstrates the advantages of thermal sensor over digital IR night vision cameras in the midst of fog showing that the thermal sensors can detect the heat signatures, whereas, night vision cannot penetrate through fog.

References

[1] Bassem R. Mahafza, Radar Systems Analysis and Design Using MATLAB, Chapman & Hall/CRC, 2016.

[2] Ba-Ngu Vo, Mahendra Mallick, Yaakov Bar-Shalom, Stefano Coraluppi, Richard Osborne, III, Ronald Mahler, Ba-Tuong Vo, "Multitarget Tracking", Wiley Encyclopedia of Electrical and Electronics Engineering, Wiley, Sept. 2015.

[3] Kursad Yildiz, "Electronic Attack and Sensor Fusion Techniques for Boost-Phase Defense against Multiple Ballistic Threat Missiles", Master Thesis, Naval Postgraduate School Monterey, 2005.

[4] CommandSpaceCereberus Datasheet, www.FLIR.com, FLIR systems Inc., 2015

[5] Sergio Miranda, Chris Baker, and Karl Woodbridge, "Knowledge-based resource management for multifunction radar", IEEE Signal Processing Magazine, 23(1), 66-76, 2006.

[6] Joseph R. Guerci, Next Generation Intelligent Radar. 2007 IEEE Radar Conference, 2007, 24(1):7~10.

[7] Ji Zhang, Yu Liu, "Single Maneuvering Target Tracking in Clutter Based on Multiple Model Algorithm with Gaussian Mixture Reduction", Journal of Applied Research and Technology, Vol. 11, October 2013.
[8] David Kondru, Mohan Krishna., "Kalman Filter based Target Tracking for Track While Scan Data Processing", 2nd IEEE International Conference on Electronics and Communication Systems (ICECS), 2015, Pages: 878 – 883.
[9] Binbin Cheng, Research on Bionic Processing for Auto-Adaptive Radar Waveform, Doctor Thesis, Tsinghua University, 2009.
[10] Cheng Bin-bin, Zhang Hai, Zhang Xiaoping and Li Hesheng, "Bats' Acoustic Detection System and Echolocation Bionics", 2012 IEEE Radar Conference,08 June 2012, Atlanta, GA, USA.
[11] Corcoran AJ, Wagner RD, Conner WE (2013) Optimal Predator Risk Assessment by the Sonar-Jamming Arctiine Moth Bertholdiatrigona. PLoS ONE 8(5):e63609. doi: 10.1371/journal.pone.0063609.
[12] Paul Zarchan., Tactical and Strategic Missile Guidance, AIAA, Vol.239, 2012.
[13] Andeggs, "3D Spherical Coordinates", https://en.wikipedia.org/wiki/Spherical_coordinate_system#/media/File: 3D_Spherical.svg, 4th Aug 2009.
[14] Mohinder S. Grewal and Angus P. Andrews, Kalman filtering theory and practice using MATLAB, 2nd Edition, 2001 Wiley & Sons, Inc
[15] Robert M. Rogers, Applied Mathematics in Integrated Navigation Systems, AIAA Education Series, July 2000.
[16] Weisstein, Eric W. "Helix." Math World-A Wolfram Web Resource. http://mathworld.wolfram.com/Helix.html
[17] Nahum Shimkin, Estimation and Identification in Dynamical Systems, Lecture Notes, Fall 2009, Department of Electrical Engineering, Technion-Israel Institute of Technology
[18] Pierre Minvielle, "Decades of Improvements in Re-entry Ballistic Vehicle Tracking", IEEE Aerospace and Electronic Systems Magazine, Volume 20, Number 8, August 2005.
[19] David Kondru, Mohan Krishna., "Kalman Filter based Target Tracking for Track While Scan Data Processing", 2nd IEEE International Conference on Electronics and Communication Systems (ICECS), 2015 Pages: 878 – 883.
[20] David C. Swanson, Signal Processing for Intelligent Sensor Systems with MATLAB®, pp. 95-101, Second Edition, CRC Press, 2011 Taylor and Francis Group.
[21] User's Manual, Digital Camcorder with IR Night Vision, WEILIANTE Co., Ltd.
[22] Compact Series User Manual, Seek Thermal, https://www.thermal.com/uploads/1/0/1/3/101388544/compact_series_user_manual-web.pdf

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 83-90 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

LabVIEW Development for an Intelligent Management System of the Electrical Energy Free Market

Rodrigo Reina Muñoz*, Rafael Valim Xavier de Souza, Alex Soto da Silva

Center for Engineering, Modeling and Applied Social Sciences (CECS), Federal University of ABC - UFABC, ZIP Code, 09210-180, Santo André, Brazil

ARTICLE INFO

Article history:
Received: 05 January, 2019
Accepted: 28 February, 2019
Online: 15 March, 2019

Keywords:
Demand response
Intelligent management
Energy free market
LabVIEW

ABSTRACT

Consumers of electrical energy have looked more closely to the energy consumption in last years due to the successive increase of the electricity bill. Nowadays it is possible to notice that many consumers have been looking for ways to reduce their consumption or the value of their electricity bill. Such a cost reduction is obtained mainly by means of own generation, distributed generation, investment in electrical efficiency, migration to the open market, or intelligent demand management. Considering that there are opportunities to reduce the bill at the end of the month for consumers that opt for the Free Market, particularly companies that has the possibility to reduce their energy consumption, is proposed here the development of a system (hardware and software) implemented in LabVIEW and hardware resources of National Instruments company, to make an intelligent demand management system for a free consumer. The program looks optimize the financial results through the demand response concept taking advantage of the price difference in the spot market. A study was done for the Federal University of ABC - UFABC in Santo André Brazilian city looking for check the implemented system. Authors hope that the results obtained with this system can help companies and consulting firms in the decision-making process as some companies can eventually adopt the energy open market. Therefore, this article seeks contribute to the reduction of monthly costs of companies through the developed system. The importance of this system lies in the fact that it is possible to identify loads considered as non-priority and can therefore be temporarily disconnected in order to maintain energy consumption within the limits established in contracts with the energy concessionaire. The results presented here are quite promising taking into account that the intelligent management of demand in the free energy market is still underexplored.

1. Introduction

Nowadays, we see a race for energy consumption to be efficient and optimized, with companies in general (public or private sector) seeking the most diverse opportunities to reduce energy bills. In this context, there is an opportunity for some consumers to migrate to the free energy market. Migration allows a reduction in energy cost. However, it is possible to go further through intelligent management of energy demand. The migration to the free market causes the contracted energy price to be lower than the price practiced by the local distributor. However, within the energy market itself, there are times when the spot price (short-term price) is higher or lower than the contract price.

Through intelligent consumption management, it is possible to further reduce the value of the energy bill if demand responds to these prices. For these reasons, this work is inserted in an extremely current context. This work will present analyzes for a university campus in Brazil seeking to reduce its cost of electric energy through the intelligent and automatic management of demand, taking advantage of price changes in the free energy market to reduce consumption at the best times, so that the consumer not only benefits from the reduction itself but also receives a financial value for the reduction obtained. Therefore, a LabVIEW program for free energy consumers will be developed.

*Rodrigo Reina Muñoz, Center of Engineering and Applied Social Sciences, Federal University of ABC - UFABC, Santo Andre – SP, 09210-580, *Brazil*, Cel. 55 11 982422337, rodrigo.munoz@ufabc.edu.br

In the energy market, reference is made to the amount of energy as *MWmédio* (MWaverage). Generally, the contracts entered into in the free market have the energy amount in *MWmédio*. *MWmédio* is the volume in MWh divided by the hours of the analyzed period. That is, the average consumption per hour. Thus, a consumption of 1 *MWmédio*, for example in January (31 days * 24 hours = 744 hours), means 744 MWh.

1.1. The Free Energy Market

The Brazilian energy market is currently divided into the regulated contracting environment (ACR) and free contracting environment (ACL). Each of these environments are distinct models of the market that have certain rules foreseen by law. In ACR, the consumer pays for their energy consumption based on the energy tariff of their local distributor. The distributor, in turn, carries out the hiring at auctions held by the Electric Energy Trading Chamber (CCEE), operations that are formalized by regulated contracts, and only passes on the price of the contracts to its captive consumers. In this hiring environment, the consumer has no way to improve or manage the price of his contracted energy. In the ACR environment, consumers have as compulsory energy suppliers the distributor, at a regulated tariff, and are called captive consumers [1]. In the ACL, the consumer can contract his own energy from any generator or marketer of energy in the country through bilaterally negotiated contracts, being able to negotiate his own price and buy his energy in the moments in which the prices practiced are more attractive. Consumers operating in this market by purchasing electricity in freely negotiated bilateral contracts are called free consumers [2].

Part of the power generators can sell energy in both environments, and energy marketers can only sell energy in the ACL [2,3]. The main motivation for a captive consumer to become free consumer is the immediate cost reduction after migration from one environment to another. However, ACL is still restricted to large consumers. Only consumers with a demand contracted with the distributor over 500 kW can migrate to the free market. A key difference between the ACL and ACR for the consumer is that in the ACL is directly subject to the impacts of a spot price. This spot price is called the Spot Settlement Price (PLD) and is calculated weekly before the beginning of the operating week of the system. From the scheduling of the dispatch of the electric power plants the PLD is calculated, which corresponds to the marginal cost of operation of the system (CMO). This price targets the differences between generation (or consumption) and spot market contracts. Therefore, the PLD is calculated by evaluating the current and future state of the water stored in the reservoirs, the fuel cost of the thermoelectric plants, the optimal proportion of hydraulic generation, thermal and the interchange between subsystems considering an acceptable risk for the system.

1.2. Motivation

The reduction of energy costs to a consumer after the migration is immediate. In other words, the free market by itself provides financial savings for its consumers. However, there are still opportunities not used by consumers, such as intelligent demand management, taking advantage of the hourly variation of the PLD to maximize financial results. The present study was based on a Campus of the city of Santo André, state of São Paulo, Brazil. The algorithm developed in LabVIEW will be used to gauge the gain that could be obtained in the Campus if the migration to the ACL environment is done. It is important to mention that currently free market models are used in various parts of the world, such as in various markets in the United States, Canada, European Economic Community, Latin America, and in Asia, countries such as Australia, New Zealand, China and Japan, has adopted this model, and each market has its own peculiarities [4].

1.3. Objective

The objective of this work is to develop a LabVIEW algorithm and a system that allows to make the intelligent management of energy of a consumer in order to assist in decision making within the scope of demand response in the ACL. Intelligent demand management in this context is still little explored in the way it is proposed in this work. Therefore, we propose an experimental apparatus that was assembled using a National Instruments CompactDAQ as well as suitable interface modules that operate in connection with such hardware. The results of this experiment may help companies in decision making related to the management of loads (on or off loads), since the developed algorithm indicates when it is financially feasible to disconnect a load set to sell contracted energy surplus in the ACL. It is necessary to point out that the present study does not take into account the energy demand profile of the companies, but even so, the algorithm assists in determining the economy from the appropriate shutdown of loads at certain times of the day. The results of this work can be used by companies together with specialized consultancies for better decision making. The consulting firms themselves could use the system as presented here.

Main contribution of the proposed system is to help the energy customers to obtain costs reduction of energy bill. The system allows identify load as loads with non-priority so them can be temporarily disconnected in order to maintain energy consumption within the limits established in contracts with the energy concessionaire. This initiative is important as the demand management for customers of the energy free market is still little explored.

The paper is organized as follows: Section II presents the general concept of demand response; Section III explain the concept of demand response to free consumers within the Brazilian energy market; Section IV presents the demand response program developed in this paper; in section V the results of this work are discussed; in section VI the main conclusions are drawn.

2. Demand Response Concept

The demand response takes place when the consumer makes an adjustment in their consumption profile in response to some stimulus. That is, the demand response can be defined as the payment of incentives designed to induce a lower consumption of electricity at times when prices are high or when system reliability is impaired. A demand response policy can result in three general types of consumer responses: The first, known as load shedding, occurs by reducing electricity consumption during peak periods of the system, when prices are higher, without the consumption pattern of other periods is changed. This measure results in a loss of momentary comfort for the consumer. An example of this type of response is when the thermostat settings of heaters or air conditioners are changed temporarily [5]. A second type of

response, called load shifting, is when consumers respond to high prices by shifting consumption at peak times to off peak times. An example would be the change of some residential activities (e.g. washing machines, dryers or pool pumps) to off-peak hours [5]. The third form of response is when the consumer uses his own generation, in which case the consumer must have distributed generation. From local generation consumers get little, if any, change in the pattern of electric power consumption. However, the generation will be providing more capacity for the system, since the system sees that the consumer has reduced its consumption of electrical energy [5].

The demand curve of consumers who are not subject to any demand response program (DR) is a vertical straight line. That is, demand is inelastic. When consumers somehow feel the short-term price changes, the demand curve becomes a straight inclined line, that is, demand is elastic. The different demand curves can be observed in Figure 1 [6].

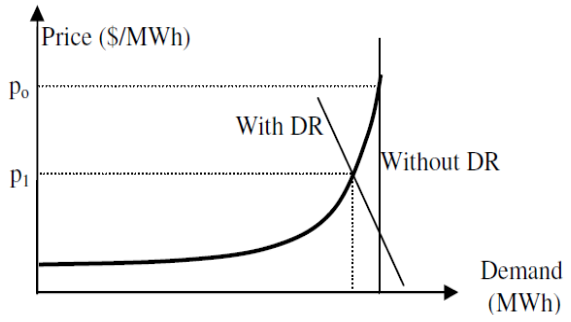


Figure 1. Simplified effect of DR on short-term electricity prices [6].

The demand response causes investments in new transmission lines and generation units to be postponed. When demand response programs act, the lines are not overloaded. In other words, the program provides more capacity for the system [7,8].

3. Demand Response Applied to the ACL

The Brazilian free energy market has a natural incentive of demand response due to the use of the PLD to compare the value differences established in the contract and consumption. The demand response in this context can be obtained through the load shedding or load shifting models. Load shedding is usually observed in commercial consumers and load shifting in industrial consumers.

In practical terms, when the consumer has more energy contracted than consumption for a certain hour of the day, it is said that the consumer has sufficiency of energy contracts at that time, which is valued to the PLD and credited to the consumer. If the consumer has less energy contracted in relation to the consumption of energy at a given time, it is said that this consumer is lacking in energy contracts, in which case the difference will be valued to the PLD and the amount must be paid by the consumer. All accounting and financial operations involving this methodology are the responsibility of the CCEE.

The mechanism of hourly determination of contractual differences means that a particular consumer can choose to reduce its consumption so that in a purposeful way might exists excess of energy as established in contract, resulting in a balance receivable by CCEE in order to reduce its energy bill power. In order this action might be successful, the consumer must map which loads are non-priority and what are the impacts of disconnecting those loads at certain times of the day.

4. Demand Management Program

Based on the aforementioned mechanism, this work proposes the implementation of a demand response program that seeks to optimize the financial result of a free consumer by cutting non-priority loads subject to a criterion that is a function of the energy contract price in the market free and spot price. There are two main parameters that should be used as a reference in a demand response program for a free consumer. These are: contracted demand with the distributor (DC) and price of energy contracted in the free market (PC). Based on these two factors and on pre-established criterion the program must make the decision to cut or not part of the load at a given instant of time. The criterion that can be pre-established refer to the definition of the loads that can be considered non-priority for the operation of the company on certain days when it is necessary to cut certain loads, and thus, be able to save energy considering the valuation in the PLD.

In the case of the demand contracted with the distributor, the decision is made based on the rule established by the National Electric Energy Agency (ANEEL) that if the measured instantaneous demand exceeds 5% of the contracted demand, that consumer will pay a higher tariff called of exceeding demand. Thus, if the program identifies an increase of more than 5% in DC, there should be a cut in the load (some loads must be temporarily disconnected) so that the consumer is not penalized in relation to the PC, criteria should be established observing the behavior of the short-term market energy price (PMC) to determine if a load-cutting action will be required. If the PC is higher than the PMC, there will be no need to cut the load. However, when the PMC is higher than the PC at any given time, part of the load must be cut off to have the power margin to be consumed in accordance with the contract. If the energy saved is large, there may be a margin of energy to be settled in the short-term market, allowing a positive financial result to be credited to the consumer included in the demand response program. The definition of non-priority load sets depends on each consumer. In the case of a university, for example, it is possible to list as non-priority and cuttable loads: set of elevators, escalators, set of air conditioners, part of the lighting of common areas, etc. The concept used to define non-priority loads is that they are loads that can be cut without generating activity and discomfort. Figure 2 shows a flowchart of a consumer for which, within a finite set of **N** loads, three sets of loads were defined as capable of acting in the demand response program. In this specific case, the cut-off criterion for the PMC was defined as follows, and may vary from consumer to consumer:

If PMC ≤ PC: there is no load cut;

If PC < PMC ≤ PC + 25%: there is a cut off of load 1;

If PC + 25% < PMC ≤ PC + 50%: there is a cut off of loads 1 e 2;

If PMC > PC + 50%: there is a cut off of loads 1, 2, 3.

Since the order of priority is:

Load N > Load N-1 > … > Load 3 > Load 2 > Load 1

That is, the first N-1 loads are of high priority and cannot be switched off.

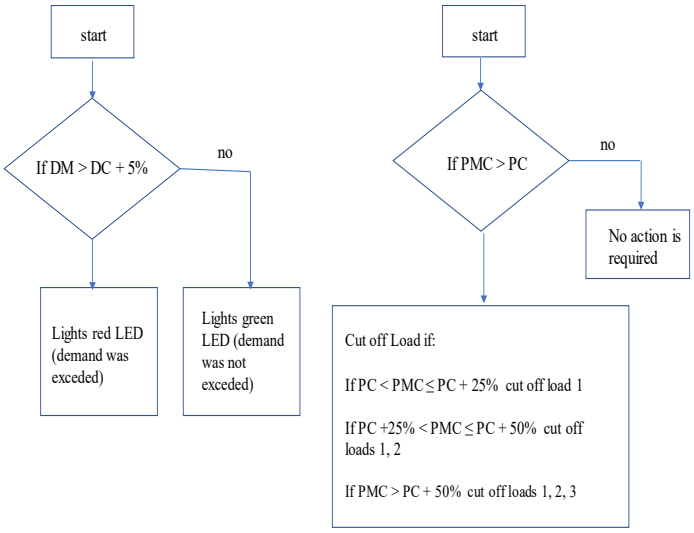


Figure 2. Flowchart for demand management program implementation [5].

The criteria for cutting the load must be defined in a specific way for each consumer according to their load pattern. To determine the economy of the consumer that participates in the demand response program, the calculation methodology, defined in (1) [9], was used. That is, the financial economy is the sum of the period of 1 to ***n*** hours of the difference between PMC and PC, when there is the intervention of the demand management program, multiplied by the demand [9].

$$VE = \sum_{i=1}^{n} \frac{(PMC - PC) * DR}{1000} \quad (1)$$

Where:

VE is the economy due to the demand response in a month basis, in R$;

PMC is the price of energy in the short-term market, in R$/MWh;

PC is the contracted energy price, in R$/MWh;

DR s the reduced demand due to the demand response action, in kW;

n corresponds with the hours of the month.

When PC > PMC, there is no load cut and DR will be zero, not contributing to the sum. The value of *n* varies according to the number of days of the months of a given period considered in the analysis. In the case of Brazil, in the Southeast region, the value of n is 673 for a month of 28 days with the end of summer (February), 720 for a month, considering months of 30 days (April, June, September and November), 744 for a 31-day month (January, March, May, July and August) and 743 for a 31-day month with start of summer season (October). The calculation of the number of hours is determined by multiplying the number of days of the month by 24h/day, and at the begin of the summer season must be subtracted one hour in the month, and at the end of the summer season one hour in the month is added.

It is also possible to determine the hourly economy, in R$, from the action of a given load cut off by using (2) [9].

$$VE_{horaN} = \frac{(PMC_{horaN} - PC) * DR_{horaN}}{1000} \quad (2)$$

Where:

VEhoraN is the economy due to the demand response in a month, in R$ during an hour N;

PMChoraN is the price of energy in the short-term market, in R$/MWh during an hour N;

PC is the contracted energy price, in R$/MWh;

DRhoraN is the reduced demand due to the demand response action, in kW during an hour N;

n are the hours of the month.

4.1. LabVIEW Implementation

Virtual instrumentation is a powerful tool for automation designers using conventional computers and specific hardware instead of traditional instrumentation. LabVIEW is software that allows the user to design programs for instrumentation, automation and process control, using a friendly interface. The CompactDAQ platform was used in this project and basically consists of a console that works in conjunction with easily configurable analog/digital input and output modules [10]. Spot energy prices are published weekly in a specific CCEE worksheet, prices that are inserted in an algorithm elaborated in LabVIEW, so that, from the date informed by the computer is able to find the corresponding spot price in the spreadsheet, providing to the management program demand updated values constantly. Details of the implementation of the algorithm can be seen in [9,11]. Figure 3 shows the spreadsheet provided by CCEE in Excel (CSV format), and Figure 4 shows the worksheet already incorporated in the LabVIEW environment which is used for calculations made by the elaborated algorithm. The import of the pricing information released by CCEE is performed through a routine that identifies the month, start date and end date information of each of the lines of the CSV file. This information is read by the program and stored in an array as shown in figure 4.

For the main demand management routine, the data of DC, PC, PCM and measured demand (DM) are used. The DC and PC variables are obtained through the signed agreement and must be informed by the user. The PCM variable is obtained through the spreadsheet provided by CCEE and the spot price verification algorithm. However, to obtain the DM it is necessary to make the acquisition of data of the demand, being necessary the use of an acquisition module of specific data for such function. That is, it is necessary to make the voltage and current measurement, or make the direct measurement through a wattmeter to inform the measured demand to the demand manager.

With the variables in hand, the algorithm presented in the flowchart of Figure 2 is implemented. The DM variable is compared to the DC variable and if DM is 5% greater than DC, a time interval is triggered to ensure that the demand is exceeded. If there is no such confirmation time, there may be undue load cuts as there may be increased demand for a short time due to activation of heavy loads. After the confirmation time is counted, the program checks the PMC (short-term market energy price) and the PC (free market contracted energy price).

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Mounth	Week	Start Data	Final Data	Heavy	Medium	Light	Heavy	Medium	Light
2	2017	9	1	26/08/2017	01/09/2017	452,85	452,85	442,35	452,85	452,85	442,35
3	2017	9	2	02/09/2017	08/09/2017	499,38	499,38	488,05	499,38	499,38	488,05
4	2017	9	3	09/09/2017	15/09/2017	533,82	533,82	533,82	533,82	533,82	533,82
5	2017	9	4	16/09/2017	22/09/2017	533,82	533,82	533,82	533,82	533,82	533,82
6	2017	9	5	23/09/2017	29/09/2017	533,82	533,82	533,82	533,82	533,82	533,82
7	2017	10	1	30/09/2017	06/10/2017	533,82	533,82	533,82	533,82	533,82	533,82
8	2017	10	2	07/10/2017	13/10/2017	533,82	533,82	533,82	533,82	533,82	533,82

Figure 3. PLD sheet released by CCEE [5].

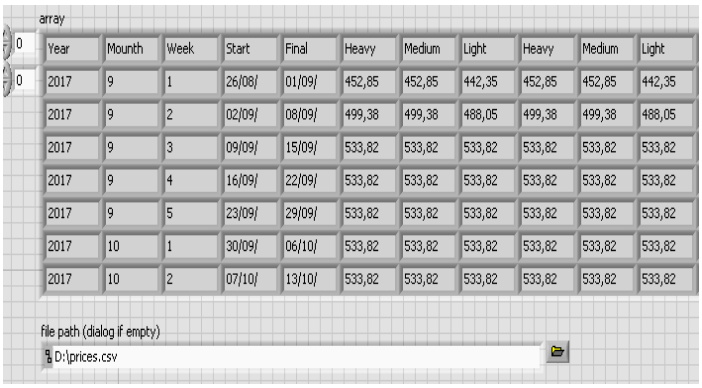


Figure 4. Importation of Short-Term Price Information (PLD).

According to the flowchart conditions of Figure 2, the result of the comparison can lead to turn off one, two, or three loads. To facilitate the implementation of this algorithm was used the MathScript tool, which allows the elaboration of programming in a similar way to Matlab. To cut only one load, the established condition for this is that PCM is between PC and PC + 25%. If the condition, however, indicates that the PCM is between PC + 25 and PC + 50%, two loads will be turned off. Three loads are cut off when the PCM value is above PC + 50%. It is important to emphasize that the criteria used for the disconnection of loads must be studied and must be according to the needs of the customer, such as the loads that must be disconnected and the adopted criteria of values between PC and PCM.

With the load-cutting information provided after processing in MathScript, the load-cutting information is decoded to be used as a digital signal. The cutoff indication occurs through signaling in the program itself and through digital signals through a specific module, and in this algorithm, each digital output represents a load. That is, when a digital signal is triggered, it indicates the cut of that load. Figure 5 shows the demand management program, while Figure 6 illustrates the experimental apparatus assembled for conducting the tests.

The apparatus of Figure 6 consists of obtaining instant consumer demand measurement through the CompactDAQ for processing in LabVIEW according to the algorithm implemented. Based on the algorithm, it will be determined whether any load set should be cut. In case of cut-off, the load-disconnection operation takes place through CompactDAQ.

In the case of load cut due to exceeded demand, a delay time (Elapsed Time in figure 7) was inserted in the algorithm to ensure that the loads are cut only if the demand is at least 10 seconds above the contracted demand plus 5% (Figure 7). That is, the LabVIEW program identifies the ultrapassing of load, but will act to cut the load only if this condition is maintained for 10 seconds.

The purpose of this mechanism is to ensure that instant demand peaks, such as for example, when a motor starts, will cause the demand manager to act.

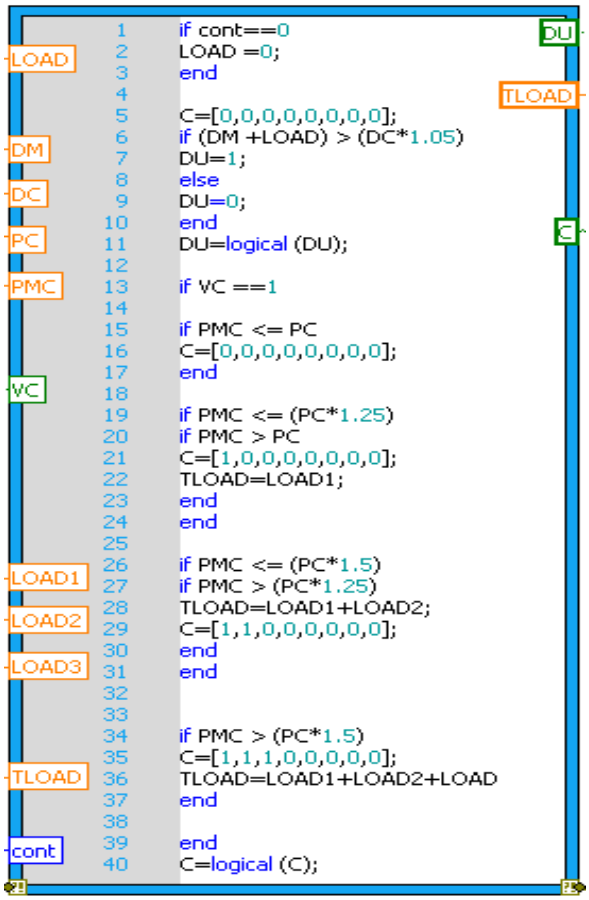


Figure 5. LabVIEW Demand Management Program [9].

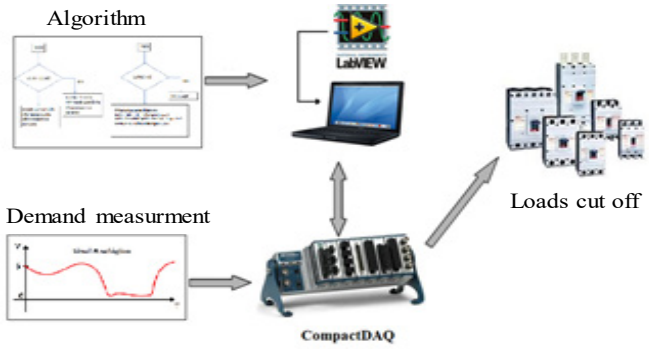


Figure 6. Illustrative pilot set-up to implement the Demand Management Program [9].

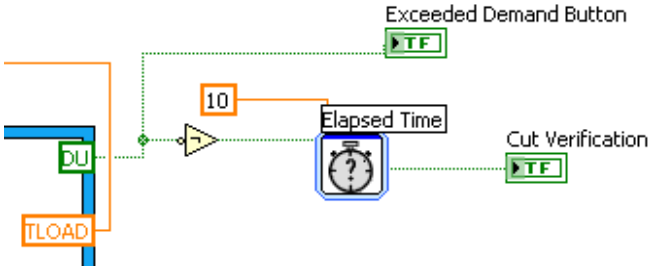


Figure 7. Delay time for load cutting action.

Also included in the program were the hourly savings in R$, savings that comes from the action of that load cut in accordance with (2) as described above. This is illustrated in figure 8.

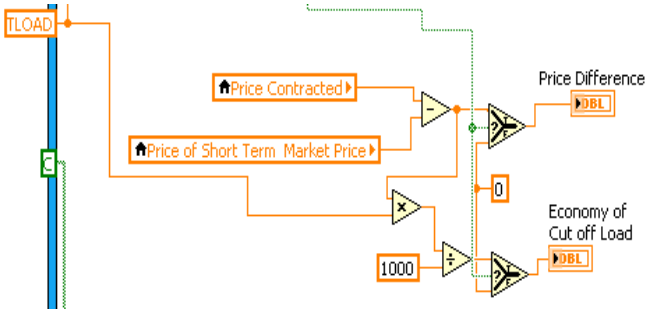


Figure 8. Hourly economy calculation included in the algorithm.

The complete LabVIEW program integrating all parts is illustrated in Figure 9.

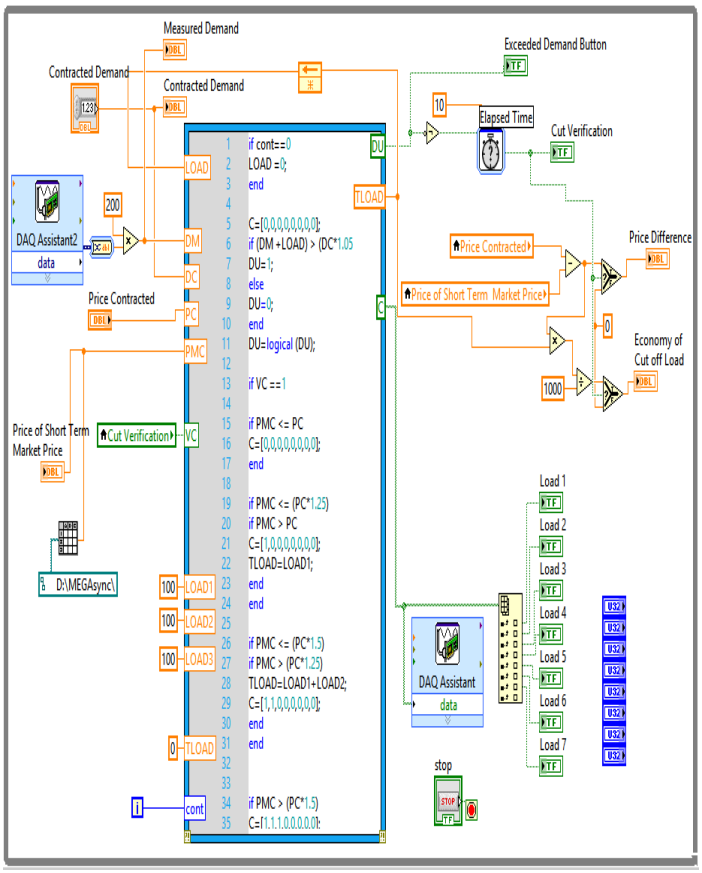


Figure 9. Complete LabVIEW algorithm of the demand management program.

4.2. Graphical Interface

In the graphical interface environment of LabVIEW an initial graphical interface was developed, shown in Figure 10, with the information of Measured Demand (kW), Contracted Demand (kW), Contract Value (R$/MWh), Short Term Price R$/MWh) and Hourly Economy (R$/h) obtained with the demand response manager. All these items are presented in the program in a simple numeric display.

In this graphical interface was inserted a button that will indicate when the instantaneous demand exceeds 5% of the value of the contracted demand, since after this limit of tolerance the consumer will pay a fine to exceed the demand. This criterion is valid for all types of consumers, whether captive or free, within or outside a demand response program. The aim of this indicator is to alert the consumer to the fine for exceeding demand, thus avoiding financial losses.

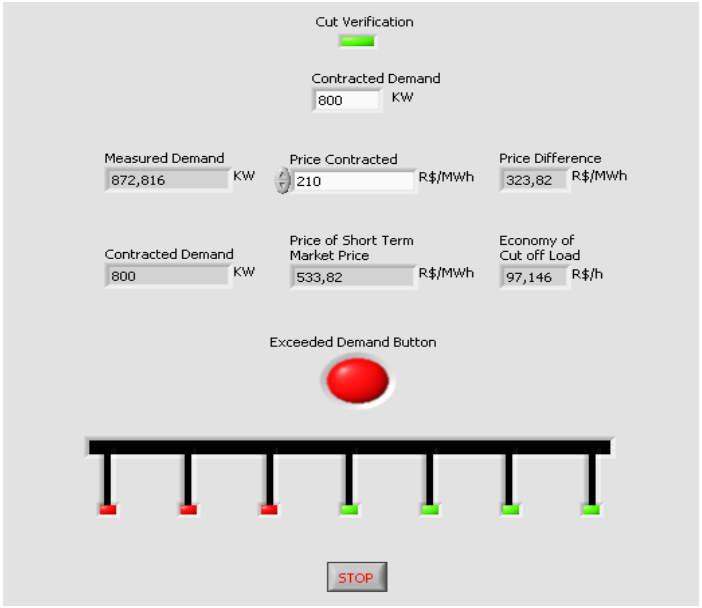


Figure 10. Graphical demand manager interface in LabVIEW [9].

Still in the graphic interface elaborated (Figure 10), there is a representation of the set of seven loads, which will be cut or re-connected, according to their order of priority. In the figure, the loads are represented by LEDs, with green LEDs representing active loads (priority), and in the red color representing loads that must be switched off in accordance with the established load-cutting criteria.

5. Test and Results

An experimental apparatus was developed that simulates the physical action of the demand response program for a consumer. As previously mentioned, a CompactDAQ was used as well as the NI-9219 and NI-9401 modules used to obtain the input and output variables, respectively, as shown in Figure 11. In this system, the input voltage on the NI module 9219, for verification purposes, represents the measured instantaneous demand. The voltage is measured and a constant conversion factor was applied to convert this information into demand values in kW.

In a practical situation, this information would be provided to the demand manager by the CT (current transformer) and TP (power transformer) of the measuring cabin equipment. A set of LEDs were connected to the output module, where each LED, as already mentioned, represents a load set. In the experiment, the LEDs lit (green LEDs) to indicate that the charging set is on. When the program sends the signal to turn off a certain LED, in practice, this action would correspond to the shutdown of that load. Figure 12 shows in more detail the highlight number 2 of figure 11, where the LEDs indicate the loads that must remain connected and the loads that must be disconnected according to the management algorithm.

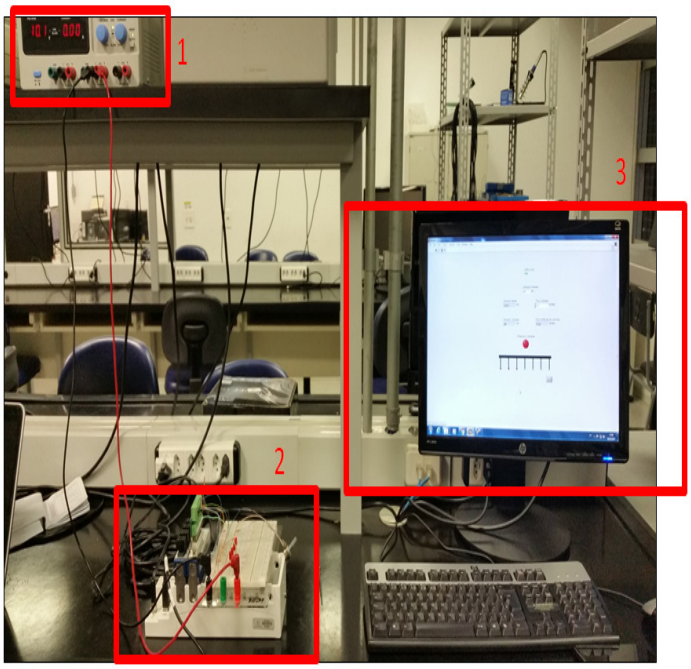


Figure 11. Experimental set-up used.

The results using the CompactDAQ were effective since, given an input signal corresponding to the measured demand, the CompactDAQ assembly was able to operate through a digital output signal that determined which output points should remain on or off. With respect to the input signal, in the experiment represented by a 10 V DC voltage source, it was obtained through a voltage divider with a potentiometer used to change the voltage.

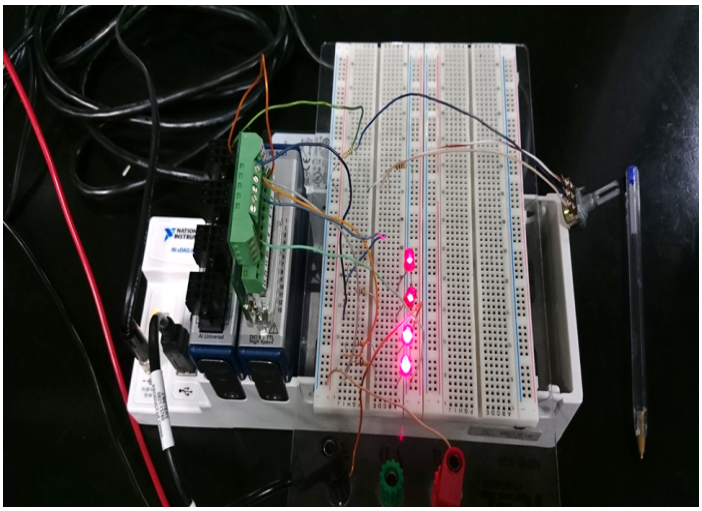


Figure 12. Experimental Circuit in operation.

In a practical system it would be equivalent to the output of the TC or TP of the meter, with corresponding corrections of that meter's constants. The output module used provides the digital signal of which ports should be switched on or off. In practice, this would be equivalent to signaling which relays/circuit breakers should be on or off at any given time, and each load set should be in a different electrical installation. To complement the experiment and identify quantitatively what could be the economy provided by the proposed demand management program, we verified the results that could be obtained at the Santo André campus of UFABC. Tests were carried out considering the period from January to September 2017, considering that contracted campus demand is 1.400 kW, energy contract price of R$ 210.00/MWh and considering the cut criterion presented in the methodology of this work. Table 1 shows the results obtained.

Table 1. Results obtained with the Demand Response algorithm

Months	*Monthly Average PLD (R$/MWh)*	*Total Economy (R$)*
janeiro-17	121,44	0,00
fevereiro-17	128,43	0,00
março-17	216,24	1.201,87
abril-17	371,47	35.976,24
maio-17	411,49	49.873,24
junho-17	124,7	0,00
julho-17	280,81	11.977,75
agosto-17	505,95	67.049,14
setembro-17	521,83	67.626,36
	Total (R$)	**233.704,60**

By the analysis of table 1, it is possible to observe that the program of Demand Response did not result in improvement of results in the months of January, February and June. These months were times when the energy price in the short term (PLD) was lower than the contract price (R$ 210.0/MWh). The results are significant for all those months in which the PLD is higher than the energy contract price (R$ 210/MWh). In the analyzed period, from January to September, a potential savings of R$ 233.704,60 from the proposed Demand Response Program was observed, according to the values presented in table 1.

The low PLD moments happen when the electrical system is in favorable conjuncture conditions, for example, when the levels of reservoirs are within the one planned by the National System Operator (ONS) and the rains are above the historical average. When the electrical system is in unfavorable condition of rainfall and reservoir levels, the PLD rises significantly and it is at that moment that the Demand Response Program becomes interesting and creates opportunities for the consumer to reduce the energy bill with the cut load and sale of excess of energy contracted. The savings opportunities increase as the average monthly PLD increases.

6. Conclusions

A tool for intelligent demand management was developed with the purpose of using the Demand Response concept to reduce a consumer's electricity bill in the free market. Tests were applied to the Santo André campus of UFABC. The program implementation was performed in LabVIEW and the tests were performed using the CompactDAQ hardware. The tests were enough to verify all the operation of the program in LabVIEW. Authors believe that the system developed is a novelty. The system allows the temporary disconnection of non-priority loads contributing to

reduce energy consumption and consequently reducing the energy bill of the users of the free energy market.

It was verified how much the campus could economize with the demand response program based on the months of January to September 2017. As results for this period, a reduction potential of R$ 233.704,60 was observed, referring only to the demand response implemented through the Demand Management Program. It is important to emphasize that the results obtained are preliminary, and, therefore, more tests are required with the application of the developed tool. It is also important to highlight that the scenario is an example of the use of the system developed to verify both the software and the implemented hardware. Thus, the loads chosen by the authors as non-priority loads were disconnected, and, therefore, the results are representative of this scenario. Thus, the key conclusion of this paper in light of the results obtained is that, Intelligent Demand Management, by observing hourly price variations can result in an excellent opportunity to reduce free energy consumer costs and relief to the electrical system. It is hoped, therefore, that the system developed in LabVIEW can assist companies and consultants with regard to decision-making in regard to obtaining savings by migrating to the ACL.

Acknowledgments

The authors thank the Federal University of ABC for the support given to this work.

References

[1] CCEE – Câmara de Comercialização de Energia Elétrica, “General operations overview at CCEE (In Portuguese)”, São Paulo, 2013, in: www.ccee.org.br/StaticFile/Visao_Geral_das_Operacoes_CCEE_2013.pdf, accessed at: 09/30/2017.

[2] ABRACEEL – Brazilian Association of Electric Energy Traders, “Booklet: Free Energy Market (In Portuguese)”, in: www.abraceel.com.br/archives/files/Abraceel_Cartilha_MercadoLivre_V9.pdf. Accessed at:10/07/2017.

[3] Rodrigues Flávia, “Energy Contracting Programming Considering Distributed Generation in the New Model of the Brazilian Electricity Sector (In Portuguese)”, PhD Thesis, UFRJ, Rio de Janeiro, 2006.

[4] Luciano Macedo Freire, “Electric Energy Markets, International Experience and the Brazilian Model of Energy Trading (In Portuguese)”, in:http://www.abraceel.com.br/zpublisher/paginas/Download_file.asp?nome=...%20%C2%B7%20Arquivo%20PDF, accessed in: 03/16/2016.

[5] CCEE – Câmara de Comercialização de Energia Elétrica, “Free Consumers in the scope of CCEE (In Portuguese)”, in: www.ccee.org.br/search/query/redirect.jsp?qid=259786&did=3193776&pos=1&idx=1&fid=&pdfq=%22consumidores%20livres%20treinamento%22, accessed at: 10/14/2017.

[6] Albadi, M. H.; El-Saadany, E. F., “Demand Response in Electricity Markets: An Overview”, Power Engineering Society General Meeting, IEEE, 2007. https://ieeexplore.ieee.org/abstract/document/4275494.

[7] Chakrabarti, B.; Bullen, D., Edwards, C., Callaghan, C., “Demand Response in the New Zealand Electricity Market”, Transmission and Distribution Conference and Exposition (T&D), IEEE PES, 2012. https://ieeexplore.ieee.org/abstract/document/6281718.

[8] Jidong, L., Zehui, Z., Li, Z., Xueshan, H., “Evaluating Short Term Benefits of Demand Response”, Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 4th International Conference on. 2011. https://ieeexplore.ieee.org/document/5994079.

[9] Souza, R. V. X., “Migration Study of the Santo André Campus of UFABC for the Free Electric Power Market and LabVIEW Development of an Intelligent Demand Manager (In Portuguese)”, Master Dissertation, Federal University of ABC - UFABC, Santo André, Brazil, 2017.

[10] NATIONAL INSTRUMENTS, “Getting Started with LabVIEW”, in: www.ni.com/pdf/manuals/373427f.pdf, accessed at: 03/12/2017.

[11] Souza, R. V. X., Silva, A. S., Munoz, R.R., “Development of an Intelligent Energy Demand Response Program for a Brazilian Free Consumer of Energy”, in: VII Brazilian Symposium on Electrical Systems - SBSE, Niteroi – RJ, 2018. https://ieeexplore.ieee.org/document/8395579.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 84-92 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Web Authentication: no Password; Listen and Touch

Viorel Lupu*

Research & Development, Online Services SRL, Buzău, 120191, Romania

ARTICLE INFO

Article history:
Received: 15 December, 2018
Accepted: 22 January, 2019
Online : 26 January, 2019

Keywords:
Web
Security
Authentication
Multi-Factor
Image-Password
Voice notification
Mobile phone
Password
Disabilities
Smart-Headphones

ABSTRACT

Just as electricity has an essential role in our lives, the internet network and especially web services have become of vital importance nowadays. Without security service layers, apparently small things like checking a child's school schedule on web may turn the daily routine into a nightmare. Web services users are still required to use many combinations of usernames and passwords. Despite technologically advances that bring many benefits to those owning top of range smartphones, complex combinations of identifiers and passwords are still required for basic security. Top tier smart mobile phones also add device-specific vulnerabilities to the risk of misuse or may expose sensitive data like biometrics. To meet users' expectations, authentication systems must be safe, fast, efficient, intuitive and easy to use, especially on mobile phones. User satisfaction, reduced fraudulent authentication issues, increased security, reduced management costs, regulatory compliance are main goals for the new advanced web technologies systems. This paper presents some real-time multi-factor authentication methods that uses voice calls to communicate random passwords to registered users. The ultimate goal is to relieve web service users from the stress of memorizing complex combinations, or copying text strings for user identifiers and passwords from paper or external memory devices like mobile phones. The new features are presented for a web service after migration from the traditional authentication system to the one with the proposed new method. This work is an extension of the paper entitled "Securing Web Accounts by Graphical Password and Voice Notification" published in 2018 IEEE International Conference on Engineering, Technology and Innovation (ICEITMC).

1. Introduction

Any authentication service must be designed, built and operated on a basis of user centric vison [1]. These design standards, associated with data protection requirements [2, 3] may imply costs, time, technologies and human resources that are beyond reach for many small companies which provide simple and useful web sites like school timetables. Users must obey the site's rules and the method of authentication. They must use given usernames and long, maybe complex or random passwords. There are several options, such as saving the login data in the browser or writing them elsewhere. If the web service is provided by an authority such as a bank or a governmental agency, the importance of a proper user access goes higher as risks of identity frauds are also high [3]. Every website that is part of our digital life, which requires and maintains personal data is equally important; without adequate security measures, any of these websites can be a weak bastion after which we might become cyber-victims.

In terms of security, when talking about Internet web services, appropriate means of authentication must be provided so that only authorized users have access to them. At the same time, Internet web services have to be easily accessible. New vulnerabilities or changes of standard requirements may occur during operations affecting the security environment completely.

Web authentication is a process that relies most on human-computer interaction. The authentication system directly influences the quality of Internet web services, in terms of usability. There is a tradeoff between the complexity of the authentication system and the service usability [4, 5].

Internet web services based on the usage of sensitive data, require high quality security systems, often a combination of SSL protocol enabled connections and multi-factor authentication systems. Low-cost single-factor authentication systems based on text strings must be reconsidered. They are no longer feasible neither for user identity nor for passwords. The RFC 2196

*Viorel Lupu, +40744545600, viorel.lupu@onlineservices.ro

recommendations have posed new restrictions on the user, as more vulnerabilities have been revealed [4, 5, 6 and 7].

On the other hand, it is difficult for the human user to remember long or complex identifiers and passwords, not to mention in stressful environments or in physically challenged contexts, in which such authentication methods are proved to be inappropriate. As a result, human user often resorts to storing identifiers and passwords in browsers memories or in easily accessible, visible files. They tend to use predictable graphic passwords and, sometimes, the same identifiers and passwords are in use for multiple websites [4, 6].

Also, we would argue that text-based authentication systems are not suitable for smartphones. Often smartphone users find it difficult to remember, read and type text during the authentication process, such as for example a simple 'copy-paste' action of a unique text passcode received by SMS, an operation which may prove to be a challenging task on a mobile phone.

Multifactor authentication is considered the best practice for authentication [4] today. Users have to provide more and more information to the authentication system in order to authenticate successfully. Authentication standards like Universal Authentication Framework (UAF) or Universal Second Factor Authentication (U2F) proposed by Fast Identity Online Alliance (FIDO) are based on latest knowledge about public key cryptography [4, 5], biometrics, mobile devices, server technologies, best practices in the field etc. in order to provide an open and public accessible framework [8].

Internet services authentication systems have to assure that secure protocols run properly: digital certificates need to be valid [5], software updates need to be at the latest version etc. System time is also very important. Web servers keep time in different ways and more or less accurate. Synchronizing the web server local time with an external time server over the Internet it is a standard practice nowadays. If the time source is manipulated or not updated, everything may go astray mostly because digital certificates, the basis for secure web communications, contain time references. Cloudflare deploys a new authenticated time service called *Roughtime*, in an effort to secure certain timekeeping services. The publicly available service is based on an open-source project of the same name that was started by Google [9, 10].

Authentication failures result in inestimable damage and there are plenty of examples: Sony Pictures Entertainment corporation (announced on December the 3rd 2014), Yahoo (on December the 15th 2016, where 3 billion accounts were compromised in a series of incidents), Equifax (on September the 7th 2017, where 148 million data sets with personal data were stolen, including social security number, with a recovery cost of $400 million dollars) or for Marriott International (announced on November the 30th 2018, where about 500 million guests, including full names, date of birth, passport information, payments, preferences etc.). [11]. It is estimated that a third of U.S. businesses have had a customer information breached in 2017, including the information needed to authenticate their customers [12]. All these events might create a spiral of more and more security incidents that would make use of this data.

Authentication failure has new dimensions such as social engineering, mobile phone thefts and most important, revealed by recent trends, SIM-Swap or cloning the GSM Subscriber Identity Module (SIM) card. Cybercriminals look for future victims using social media websites. On such websites, it seems easy to find personal details of the victim (e.g. the date of birth, the e-mail address, the mobile phone number, etc.). Finally, it is not so hard to convince the GSM operator to make an emergency transfer of the GSM number to a new SIM (in the criminal hands). Then the authentication system works as expected. But not in favor of real users [13, 14], as the real user just discovers that his/her mobile phone is dead. From a technical point of view these fraudulent actions seem beyond the scope of the authentication process.

The security system should not rely entirely on the security components of the operating system. Recent security vulnerabilities (the case of Meltdown and Spectre) leak data as encryption keys, passwords [15], security identifiers, images etc., data that should be protected by the operating system. Instead, it is disclosed without user consent. The process of solving those vulnerabilities is a long one, as it will have to deal with hardware architecture redesign [15].

Nowadays, security is a serious concern for the entire society, being the result of technology, people and policies working together. In this sense, the new GDPR have been imposed in the EU since May the 25th 2018 [2]. In Europe, "companies must alert government authorities within 72 hours of a known breach and may be fined up to 4 percent of their global revenue under data protection laws" [17].

This study proposes a different solution for real-time, multi-factor authentication: image selection [17, 18 and 19] guided by voice via phone call notifications [20, 21]. This paper presents a detailed analysis, results, related works and comments on the implementation decisions, as well as on user reactions. Our aim is to improve the authentication process with better user experience and comfort. Also, we intend to extend its usability in high stress conditions (public hospitals, police and justice departments, public administration) and for the elderly or people with disabilities.

2. Background

Forms are commonly and efficiently filed with the use of secure web forms, be it small or medium size enterprises (SMEs), corporations or government entities. In the situation analyzed in this paper, the web forms in use run on a national authority's centralized web services application due to the need for real-time sensitive medical records such as those used in the field of human organ transplants. The operators reporting the data, taking into account the strict medical specialization and the legal implications, are medics accredited for the respective field. Each medical unit enrolled in the national transplant system has at least one accredited medic on staff. During the analyzed period, there were over 60 accredited medics in the program. Reports were filed on a daily basis, but also following certain unforeseen events.

The secure web forms are filled in using computer terminals placed in the operative staff common rooms of the respective hospitals, as terminals used for patient registration. Terminal access is permitted to all the medical staff in the hospital. Thus, the transplant database system is potentially at risk of exposure to unauthorized personnel. The credentials were disseminated to high school level medical personnel, which were operating data. Many times, the Internet address, user name and password were written

on post-it note stuck on the terminal' screens, further allowing access for curious individuals to presumably protected data. Even though such post-it's were not used in all locations, there were notebooks or unprotected files containing these credentials.

The described situation does not represent an exception; it is often met in other environments, where the legal implications may be less serious. There were also security conscious operators who used their private computer terminals or mobile devices for these reports. In the effort to limit the data exposure, several technologies and methodologies were evaluated, being clear that the use of text strings as passwords is difficult to use on mobile devices [4, 5 and 22].

An attractive and convenient alternative is the use of personal mobile devices in comparison to investing in new equipment. Nowadays, the personal mobile telephone is ubiquitous, being a smart device, permanently connected to the internet and provided with an array of sensors used for increased security. High resolution video cameras, near field communication (NFC) readers, fingerprint readers, geo-location (GPS) receivers are already embeded in current smart phones, open the opportunity to use multi-factor security solutions. Starting with 2009, the smart phone market absorbed more than 173 million units, out of which almost 2 million have an embedded NFC reader (www.statista.com). NFC terminals have the ability to read NFC chips embedded in many types of supports such as implants, ID cards, labels and rings.

The NFC technology embedded into mobile devices, based on the Near Field Communication Data Exchange Format (NDEF) allows the automatic reading of a secure web site address and users identifiers. According to the NDEF specification, a URI record for the secure website using https://domain/aplication/UserId format [20] allows the user to open automatically the web application by approaching the NFC support to the smart mobile device. Thus, increasing security, user comfort, though the potential problems could be even greater in case these NFC supports are lost or stolen.

We have also analyzed the possibility of sending one-time, unique random passwords to authorized users, based on preregistered personal mobile phone numbers. The option of sending the password through Short Message System (SMS) is almost intuitive. Banks use this system for two-factor authentication for several decades. In 2005 it was demonstrated that the use of this system can be compromised since the SMS as a communication technology is quite vulnerable [19, 22]. This vulnerability is inherited from the telephony signaling system, also called SS7, developed in 1975 and in use ever since. Due to this, the SMS is no longer a recommended communication channel in authentication systems.

Cybercriminals rapidly penetrate weak security policies on mobile devices using a combination between social engineering, Internet available advanced software technologies and the use of sniffers for text strings sent on clear text communication channels. Their target is to clone SIM card in order to take over the victim's mobile telephone number, build a replica of the security environment on a malicious device for later use in hijacking all victim's Internet accounts where the phone number is registered for multifactor authentication (e.g. bank account). Methods that were unimaginable several years ago are now available to anyone who wants to dabble with them [13, 14].

Images or sounds can be an alternative to text passwords. Image-based authentication seems to be promising especially due to the fact that the human mind retains images and image associated actions better in comparison to written text strings [4, 6, 7, 17 and 23]. The replacement of the text password with a graphic one does not increase the level of security of the authentication method, being equivalent to the use of a four-digit PIN number [23].

The Interactive Voice Response (IVR) systems have been around in call centers since the 1970s and are presently used to transmit voice codes. For example, Microsoft runs an international system for software license activation using a similar IVR technology that receives and transmits the activation codes to be typed by the user.

The system analyzed in this paper implemented the use of a graphic password, which is communicated through the telephone network.

3. A new method for web authentication

The suggested authentication method aims to increase system security and user satisfaction. Taking into consideration that users easily recall images [4, 6 and 18] and image-related actions, the proposed authentication system displays images and processes user actions (as screen touch or clicks). The images are selected to be meaningful to the users. With voice guidance through a telephone call, the users have to choose between them. Processing the user's actions, the authentication system assembles an indicator and compares it with the random one-time password constructed into voice indications.

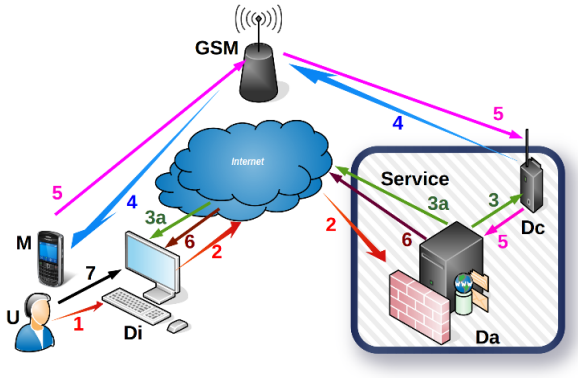


Figure 1: Web authentication system

The proposed authentication method (Figure 1) has two main phases. The first phase consists of the user's identification and starts the authentication session. The second randomly generates the password, initiates a phone call and authenticates the user. If the authentication system receives a known user identity, then the second phase is automatically started.

The first phase is of a special importance, derived from the fact that there are a lot of devices capable of initiating secure web sessions, from simple, classic personal computers to the latest mobile smart-devices with biometrics. The classic old-style computer browser provides options to save website address and user identifiers. Specially crafted computers like laptops may be equipped with additional specialized hardware to read fingerprints, NFC cards or to process images and sounds. With latest

smartphones, user identifiers may be stored and better protected e.g. by fingertips and encrypted file systems. The authentication system has to detect from the authentication request who is the potential user and from what device the authentication request is made. If that device is already registered and all secure attributes are up to date, then the authentication procedure goes forward.

The authentication process is carried out following several steps [21] as illustrates in Figure 1:

1. A secure web session (2) is established with the server system after the user's (U) action (1) upon the Internet terminal (Di). The user's action (1) is carried out either by typing the URI (Da), or by using the device's video camera visual to recognize a QR code, or entering proximity of the terminal with an NFC ID card issued by the organization running the web service, if the respective terminal has an embedded NFC reader.

2. When the SSL session is opened (2), it provides a number of details about the Internet terminal (Di) that initiated the session and about the particular user (U). If those details are read off the NFC card's chip or the browser's memory, then the system automatically continues to the next step. As such, the user has to provide an identifier related to a user recognized by the web system (Da) and a password to confirm the authentication request. The request confirmation password can be in the form of a text string, graphic image set or an audible sound, depending on the particular security policy.

3. If the secure web session is initiated with the user's identifier and a known confirmation password, the web service generates a unique one-time random password and directs (3) the telephone call initiation device (Dc) to establish a connection with the user's (U) telephone number (4) on record. Simultaneously with the telephone connection, the web system (Da) will securely update (3a) the display on the requesting terminal (Di).

4. The communication device (Dc) initiates the telephone call using the public telecom network (4) (i.e. GSM). The web service (Da) monitors the status of the telephone call (5) using the communication device (Dc). The system waits for the call to be answered before starting dictation. The telecom network can also transmit other events regarding the status of the call such as if the line is busy or the mobile phone is outside the coverage area etc. Once the user answers, the communication device (Dc) will play the voice indications appropriate for the respective unique password.

5. The user (U) answers the call or not. Events detected by the communication device (Dc) in the public telecom network are sent in real-time (5) to the web system (Da). The web system will close the web session if the telephone call fails for any reason (i.e. line is busy).

6. The web system (Da) securely communicates the status of the call with the user's terminal. Thus, when the user answers the call (5), the terminal receives and displays the set of images (6). Figure 2 is an example of a set of images displayed on the user's terminal.

Figure 2: Example of a set of images for one-time-password

5. Following the indications received in the call (5), the user selects (7) the relevant ones amongst the other images (Fig. 2) displayed on the internet terminal, thus communicating the unique password to the web servers. The system validates the password and authenticates the user if during the call the latter has selected the appropriate images in the instructed order. Afterwards, the authenticated user may continue the web session. Any deviation from these restrictions leads to the closing of the web session by the web service.

The aforementioned web authentication method has the following advantages:

The password used to start the whole process (step 1) does not directly authenticate the user. It is merely a password confirming the user's intent to authenticate into the system. If unauthorized individuals use the respective password, the process cannot be continued since they do not possess the authorized user's mobile phone terminal. The initiation of every web session generates a telephone call and in consequence, the real user receives unsolicited calls, knowing that his account is being abused. The real user can take steps and change the confirmation password. In case of malicious authentication requests, the system eliminates abuse before an imposter can access the system.

Using NFC ID cards, rings or implants, the user is relieved of the need to memorize identifiers or confirmation passwords, which can be constructed of very long strings that are practically impossible to memorize, leading to increased system security.

In the absence of NFC support on protected access devices (i.e. mobile telephone terminals with biometric sensors), the identification strings may be stored in the browser's memory or in password manager type applications. The terminal's web browser (Di) may also save the Internet address and credentials of the secure website. The website address may be displayed in the form of an icon on the display of the Internet terminal, as to ease the user's task to memorize such details. The secure web session once initiated using the icon can also provide the user's identifier and confirmation password.

If public access web terminals are used, the confirmation password can be image-based thus avoiding the possibility that the respective terminals may memorize the confirmation password in the browser or password manager applications. Still, this type of password may be visually exposed to people in close proximity or to video surveillance systems.

The images used by the user to build the password based on the telephone call instructions are displayed on the internet terminal in the step 6, only if the user answers the call. This workflow reduces the exposure of the images to unauthorized users, increasing the security of the web service.

The telephone calls can be processed by the operating system (OS) of the smartphone. Thus, we can apply policies based on time periods or in correlation to other data such as geo-fencing (GPS position), GSM network name, Wi-Fi network SSID etc. as to semi-automatically reject unwanted calls.

The communication device (Dc) can produce the sounds necessary for voice calls through synthesis (text to speech) or by combining pre-recorded sounds. The pre-recorded sounds can be the result of professional recordings. Moreover, by mixing voice with background noise, the resulting sounds can be recognized by the users as authentic web service instructions.

This authentication method is terminal OS (operating system) and web browser independent (e.g. Linux, Windows, Android,

IOS with Chrome, Firefox, Edge, Opera etc.). Suggested authentication method does not rely entirely on the operating system security components. It is crucial that the Secure Socket Layer (SSL) based on digital certificates etc. to be fully functional.

The implemented and studied system is being improved continuously. The public access terminals available to all operative hospital staff have been fitted with specialized USB connected NFC readers. With these readers and NFC ID cards, the authorized medical staff have quicker access to secured web forms used to report sensitive data, at the same time being less exposed to different security risks such as phishing attacks. As a result, the system became safer, more comfortable and much more productive.

4. Findings

The proposed web authentication service has been implemented [20] and runs continuously since December 2016 to serve the central transplant agency and authorized hospitals across one European Union member state. Beginning with September 2017 the usage of the new authentication system has become mandatory. Every authorized hospital is represented by an accredited specialist who daily reports medical events and data to the national authority, 24/7.

Web service usage data were collected continuously in log files. Data collected for this analysis is based on the regular system usage by 60 authorized operators in the timeframe between September 2017 and April 2018. In this time span, the number of operators increased. The analysis in this paper is based on all the web traffic for 20.000 successful authenticated sessions.

The proposed authentication method was implemented on short notice and without staff training, taking into account that it was virtually impossible to stop activities in hospitals only for this specific training. Thus, each operator has discovered the new authentication method when a new medical event or new data had to be reported to the central entity or when, in some cases, a lot of unsolicited phone calls from the web service (a phone-call for every authentication request) have been received. These unsolicited phone calls forced the specialist to change his/her password which as a consequence of the new authentication method become a confirmation password. Many specialists have learned to use private web sessions and not to store the user name and password in Internet browsers when using computers available to all the department staff. Subsequently, they received NFC ID cards and NFC readers. The Internet web service was re-engineered for a special non-public sign-in web page which started the secure web session and the authentication process by automatically reading the NFC link. This administrative procedure reduces web traffic in general, but did not affect the selected sessions for this analysis, because the use of the NFC ID card automatically generated a secure web session for a known user identifier (if it is still valid) to start the phone call process.

For this analysis all web site logs were saved and processed. The web server logs any web request, writing in the logs the client's IP address, protocol and the time moment when the web request was made, as well as the files that were requested on the server, the details of the browser, etc. [28]. Telephone communication logs contains telephone numbers, call events and timestamps. In this analysis all this data is correlated from web site logs, telephone communication logs and data from the web site database to filter user authentication requests and necessary details about the authentication process.

Web server logs contain a large number of unauthorized sessions because of the web services' public exposure. The current analysis does not take into account all unauthorized sessions. For any user, the authentication process model starts with the first GET (a method from the HTTP protocol [29]) web request for the web sign-in page retrieved from the web server log. These GET web requests are filtered for those requests that send valid user identifiers, accepted confirmation passwords and are continued with phone calls.

Then data is filtered and consolidated by web sessions having a unique client set of IP address, user identifier, session timeframe and phone call. In this stage, the user identity and telephone number are associated from the central database. Telephone calls log complete the analyzed authentication process model with call events. All sessions with an unsuccessful phone call are discarded i.e. sessions when the phone calls ended with busy signals, rejected calls, dial error, busy network or not answered at all, including phone calls with out of network coverage signals or if the user hanged-up before the end of the voice message. Then, searching the web server logs for session authentication failure or success events all necessary information is collected and processed. Now the entire authentication process is modelled. Resulted authentication process models are then ordered by user and timestamps. Figure 3 illustrates sample data processed by this model for one user (where real user identifier was obfuscated with *UserName* for security reasons).

As illustrated in figure 3, the user's authentication activity pattern reveals the fact that this particular user has tried to authenticate many times without success, especially because he/she was too slow to select the indicated images within the required amount of time. The first three lines of data reveal the fact that he/she has tried to authenticate three times on the 29th of September 2017 starting from 9:37 AM with no authentication success. Then, there is a new failed try after nearly two hours, same day at 11:17 AM. The next day, on the 30th of September the user has tried to authenticate two times from 9:36 AM and succeed the authentication one hour later, at 10:32 AM. On the 2nd of October, two days later, the user was able to succeed from the first try, at 9:58 AM. The IP address reveals that, starting with the 2nd of October, the web sessions are initiated from another network device.

Moment	UserId	IP	Call
29.09.2017 09:37	UserName	213.233.88.223	Notified
29.09.2017 09:38	UserName	213.233.88.223	Notified
29.09.2017 09:38	UserName	213.233.88.223	Notified
29.09.2017 11:17	UserName	213.233.88.223	Notified
30.09.2017 09:36	UserName	213.233.88.223	Notified
30.09.2017 10:32	UserName	213.233.88.223	Notified
30.09.2017 10:32	UserName	213.233.88.223	Ok
02.10.2017 09:58	UserName	86.124.60.4	Notified
02.10.2017 09:59	UserName	86.124.60.4	Ok
03.10.2017 06:44	UserName	86.124.60.4	Notified
03.10.2017 06:44	UserName	86.124.60.4	Ok
03.10.2017 10:49	UserName	86.124.60.4	Notified
03.10.2017 10:50	UserName	86.124.60.4	Notified
03.10.2017 10:50	UserName	86.124.60.4	Ok

Figure 3: Authentication process modeled from log files

Calculating the averages for all the users, the average number of failed authentications attempts is 250% higher as compared to the authenticated ones, in the first month of use. In the following months, similar to the phenomenon described in Figure 3, the average values decrease. The value is 70% in the second month of use. Notice the decrease in the following months: 45% in the third month, 10% in the fourth month and 4% in the fifth month. These average values form the graph in Figure 4, confirm the fact that in time the users learn the new authentication mechanism and become productive.

The graph in Figure 4 also contains failed authentication attempts launched from public access terminals. The data analysis shows on one hand that the public access system was exposed to unauthorized use and, on the other hand, that part of the medical staff who knew the system and accessed it without respecting the legal procedures, have finally understood that that it was no longer possible to do so. These authentication attempts have been reduced considerably by implementing the new method.

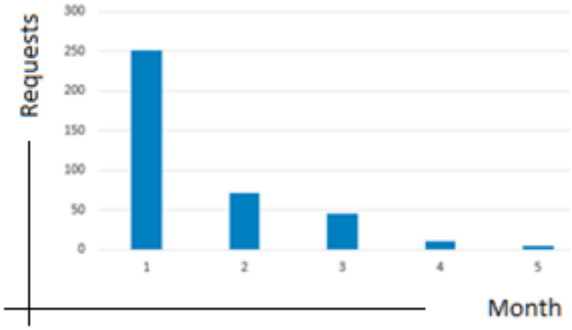


Figure 4: Learning cycle (5 months)

From analyzing the use of data, it has been revealed that users need an average of 30 seconds to authenticate, the difference being calculated between the authentication request timestamp and the start of the secure web session. The 30 second interval is composed of the time needed to place the telephone call (10 seconds) and 20 seconds needed for the user to answer the call and to select the images indicated verbally during the call. Those 30 seconds are the price paid for an increased security system and this time frame is comparable, as well as less time consuming than multifactorial authentication by other means (by text message, where the user must wait to receive the message, to open the application and must memorize the received code in order to later type it into the web interface).

While running the new system we also found that users are creative in finding new solutions to circumvent strict rules. As such, they requested that their personal telephone number to be changed in the database to the one publicly accessible in the department or have temporarily redirected their calls to other telephone numbers, of work colleagues usually. These practices are barely legal. However, we found fewer examples of such behavior as compared to the previously mentioned cases with post-it notes placed on the terminal's display in the operative medical staff room, that facilitated access to anyone who would breach the system. In order to counteract these use cases, we needed better organization frameworks, better user training and the use of advanced technologies that, for example, can detect automatic call forwarding.

One of the measures we applied, was to expedite the implementation of equipment necessary for the use of NFC technologies, such as the implementation of specialized NFC readers and user NFC ID cards. Thus, the users no longer knew the authentication web page URI, the unique user identifier nor the confirmation password. In order to initiate the session, they just needed to have the NFC ID in close proximity to the reader. The NFC ID cards improve the user comfort, they contribute to the increased system security, but, at the same time, they increase the implementation costs. The NFC ID cards were readily accepted by the medical staff, which later realized how easy they are to use.

Still, call redirection, unsecured mobile telephone theft, SIM card cloning are still major vulnerabilities for the system, especially if the voice notification subsystem uses an internationally recognized language and the images are directly indicated.

5. Related work

Since the user mentally establishes the connection between sounds and images, indirect indication is possible. For example, images can display animals while sounds describe their favorite food. In the set of animal icons, the rabbit will be chosen when the word "carrot" is heard.

The use of mother tongues can improve system security. A flower is indicated by vocalizing the expression "smells good". It is difficult for an unauthorized, non-native speaker to choose the flower icon when hearing *lukter godt* or *miroase bine* while this comes natural for a native speaker.

Indirect indications require special attention when it comes to expressing emotions. It might be harder for an unauthorized person to guess the correct images, when those images are selected in relevance to the experiences and according to the expected emotions of the actual user. Voice indications may sound cryptic because the meaning behind any symbol needs previous user initiation to understand and act in the required timeframe. An example of this kind of set of images are illustrated in figure 5.

These methods may have a difficulty level so high that unauthorized users are unable to guess correct image passwords even if they control both the authentication workstation and the user's phone. This method of authentication emphasizes the fact that the user's prior knowledge is an important element to be considered in multifactor authentication processes.

Figure 5: Example of a set of images

Security may also be enhanced within the proposed method of authentication using some traps as a result of some simple changes to the set of the displayed images or to the voice indications. This kind of changes require targeted training to user groups. This way, users will be instructed not to select any image following the first voice indication. The first voice indication in the call is a trap for users that do not know this detail. Another simple change is to use trap images. Trained users can easily avoid them even if they are urged to select them. In the image set from Figure 5, a trap image may be the icon of the key. A user with no training might touch that icon, following the voice indication and the authentication would fail.

Language that is familiar to the user (i.e. the local dialect) as well as false verbal instructions and/or trap images can provide insurmountable obstacles for unauthorized users. These methods have the advantage of being categorized as "what the user knows" [4, 5] and are not equipment-dependent (since they are compatible with fixed-line, analogue telephony) and with some necessary adjustments they can be used as an authentication method for users with severe deficiencies (such as visually impaired persons).

The web service interface must be carefully designed and implemented to hide the correlation between a trap and a failed authentication. Repeated tries must not help the unauthorized user to discover the reason why the authentication fails.

No technique is infallible and, in the domain area of security, authentication, or content protection, continuous efforts must be made to be at least one step ahead of cybercriminals. The suggested authentication method was developed on the basis that the users, the most important entity of the system, make mistakes. In order to reduce the effects of human mistakes and to help users in their overall effort to maintain the data safe, a simple security device has been added to the system. Figure 6 presents the updated authentication system; the added security device is labeled CA in the form of audio headphones.

The enhanced voice notification flow (figure 6) contains a smart-headphone (CA) as compared to the original flow (figure 1). These smart-headphones must be capable to function as regular headphones connected to the mobile phone and smart enough to detect special encoded data sub-channel in the audio stream. When such a stream is detected, the smart headphones automatically converts encoded data into comprehensible sounds according to the user's language. In other words, the headset hides a security component that users need to have in order to succeed in the authentication process.

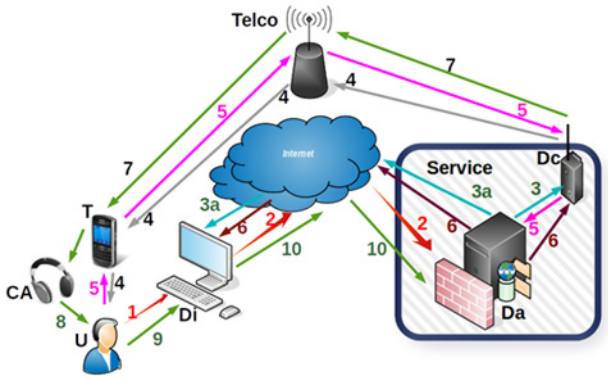


Figure 6: Enhanced web authentication system

In the enhanced authentication method, the telephone voice call contains an encrypted audio stream sub-channel, unintelligible to humans. This way the enhanced authentication method is more robust if the phone call is redirected, the mobile phone is lost, cloned etc. keeping the one-time random passwords voice indications just for certified users. One must have the smart-headphones connected to be able to listen the voice indications.

These especially designed smart headphones may be subject to the addition of more security features like biometrics or other security features that restrict smart headphone use outside of the particular scope (e.g. the user must touch the headphones to allow decryption of the audio stream, a gesture unknown to a malicious person who manages to have both the smartphone and the smart headphones, since the shape of the smart headphones and their functionality seem normal/common).

Following these ideas, a prototype of these smart headphones was made. The voice notification system has been parameterized to transmit data using well-known dual tone multi-frequency signaling (DTMF) analogue telephony technology. DTMF streams are transmitted over phone calls, simultaneously and in the same frequency band as the voice channel. DTMF audio streams are easily decoded as bit streams using low power CMOS chips (e.g. MT8870). Bit streams are transmitted to a connected low power microsystem built as a System on a Chip (SOC). SOC processes all data streams and finally produces voice indications into the headphone speakers.

The proposed enhanced authentication process flow (Fig. 6, Fig. 7) is started by the user (1) in interaction with a browser. The browser sends the user identifier, the user confirmation password and the details about the device that initiates browser request (e.g. NFC device identifier) to the server, over an SSL Internet connection. The server prepares a random one-time-password (OTP) and engages (3) the modem to call the known user and, at the same time, instructs (3a) the browser to display relevant information.

The modem tries to call the known user's phone number (4). If user's phone is busy or if there is a telephone network congestion or if the user answers or rejects the call (5) then the server either closes the web session or continues it by sending to the modem (6) a code derived from the OTP and to the browser a web page to display the set of images (6a).

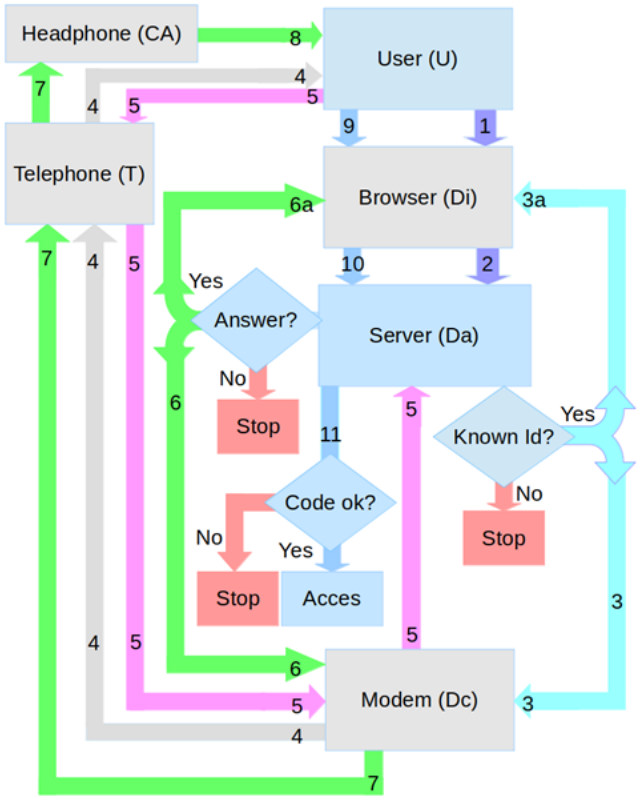


Figure 7: Enhanced authentication process flowchart

The modem starts the audio playback and mixes the DTMF encoded indications in the audio stream (7). The DTMF encoded string is detected, decoded and played back (8) by the smart headphones. The user acts (9) on displayed images based on his knowledge and according to the voice indications received through the smart headphones. The browser sends the user actions (10) (click or touch events) back to the server, in real-time. The server analyses the user web session continuously and verifies if all the required conditions are met (11) in order to authenticate the current user successfully.

Although the presented architecture (Fig. 6) contains a hidden security element in an audio headset, with virtualization technologies everything may become software based in a carefully architected mobile system. User comfort increases with virtualized devices, but the system management layer will have to address new security risks. These technological options must be thoroughly studied before being implemented as a public service.

Conclusion

As the analysis presented in this paper shows, the new authentication methods redefine the user interaction with the web system during the authentication process. These new methods create an independent channel of communication between the user and the system, through the public telephone network in order to dictate a unique, random single-use password. The presented methods remove some important security vulnerabilities such as SIM card cloning, used to take fraudulent control of web accounts.

We have analyzed in detail the operation of a medical system that implements the new manner of web authentication, presenting a series of data and use models which on one hand confirm the usefulness of these methods and on the other hand stand at the basis of future improvements. Before implementation, the analyzed medical system was plagued with unauthorized access sessions by individuals using public access terminals available in hospitals. After the implementation we have found a drastic reduction of unauthorized access coming from this type of terminals.

The central authority running the analyzed web service implemented the use of NFC ID cards and distributed specialized NFC readers for the necessary terminals. The use of NFC ID cards has become standard, noticing increased speed in the authentication process. Now, users no longer need to memorize Internet web site address, personal identifiers nor passwords in order to initiate the secured web sessions used to signal the authentication intent of the user.

Even though the paper presents a system based on random single use passwords (OTP) that are equivalent to 4-digit PIN numbers, they do not represent a restriction. The system analyzed in this paper uses such passwords because they are sufficient to obtain the desired effect, without disrupting unnecessarily the fragile balance between security, functionality and comfort.

In accordance to the presented method, the authentication process produces plenty of other useful data besides that collected during the password creation phase. The continuous analysis of the system's identification data using a Deep Learning algorithmic process could increase security as a whole by applying custom user profile policies. The user profiles contain data that refers to the timeframe needed for the voice indications to be recognized and actions to be taken for successful authentication. These ideas have not been implemented nor analyzed yet, but are considered for further development.

As the latest advances in computing (including better, more powerful processors, large and very fast storage systems e.g. solid-state drives (SSD), artificial intelligence, deep learning technologies) are accessible to the public but also to cybercriminals, further efforts must to be undertaken to enhance every component of the security system in order to fight cyber-crime. This work tries to emphasize the role of people and culture in the authentication process. The main ideas presented in this paper have been implemented / tested and analyzed in detail. As one would expect, the subject open to future research, as knowledge, innovation, technological advances in Information Technologies, communications, miniaturization and low energy consumption devices will bring new, more secure and user-friendly solutions.

Conflict of Interest

The author declares no conflict of interest.

References:

[1] ISO 9241 Ergonomics of Human-System Interaction Standard, https://www.iso.org, 2018.

[2] EU General Data Protection Directive (GDPR), https://www.eugdpr.org/, 2018

[3] Paul A. Grassi, Michael E., Garcia James L. Fenton, "Digital Identity Guidelines", NIST Special Publication 800-63-3, https://doi.org/10.6028/NIST.SP.800-63-3, 2018

[4] D. Dasgupta, A. Roy, A. Nag, "Advances in User Authentication", ISSN:2363-6149, 2017

[5] Richard E. Smith. "Authentication – From Passwords to Public Keys", ISBN: 0-201-61599-1, 2002

[6] D. Charruau, S.M. Furnell & P.S. Dowland, "PassImages : An alternative method of user authentication" in Advances in Networks, Computing and Communications 2, ISBN: 9781841021409, 2004

[7] S. Chiasson, A. Forget, R. Biddle, P.C. van Oorschot, "User interface design affects security: Patterns in click-based graphical passwords" https://link.springer.com/article/10.1007/s10207-009-0080-7, 2009

[8] FIDO Alliance, https://fidoalliance.org/, 2018

[9] Lily Hay Newman, "Clouldflare and Google Will Help Sync the Internet's Clocks - and Make You Safer", https://www.wired.com/story/clouldflare-google-roughtime-sync-clocks-security/, retrieved Nov. 2018

[10] Roughtime Project, https://roughtime.googlesource.com/roughtime, 2018

[11] Nicole P., Amie T., Adam S., "Marriott Hacking Exposes Data of Up to 500 Million Guests", The New Work Times, https://www.nytimes.com/2018/11/30/business/marriott-data-breach.html, accessed 2018-12-01

[12] State of Authentication Report 2017, FIDO Alliance, https://fidoalliance.org/, 2018

[13] "Silicon Valley Execs Targeted in 'SIM Swap' Hacking, $1 Million in Crypto Stolen", https://www.newsbtc.com/2018/11/22/silicon-valley-execs-targeted-in-sim-swap-hacking-1-million-in-crypto-stolen/, 2018

[14] "$14 Million in Cryptocurrency Allegedly Stolen By SIM Swappers, Authorities Report", https://www.cryptoglobe.com/latest/2018/09/14-million-in-cryptocurrency-allegedly-stolen-by-sim-swappers-oklahoma-city-authorities-report/, 2018

[15] Paul Kocher et.al., "Spectre Attacks: Exploiting Speculative Execution" https://spectreattack.com, 2018

[16] N. Perloth, A. Tsang, A. Santano, Marriott Hacking Exposes Data of Up to 500 Million Guests https://www.nytimes.com Nov. 30, 2018

[17] Z. Zhao, G. Ahn, J. Seo, H. Hu, "On the Security of Picture Gesture Authentication" 22nd USENIX Security Symposium ISBN: 978-1-931971-03-4, 2013

[18] Arti Bhanushali, et al., "Comparison of graphical password authentication techniques" International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 1, 2015

[19] William E. Burr et al. "Electronic authentication guideline" NIST Special Publication 800-63-2, 2013

[20] NFC Forum, "Essentials for Successful NFC Mobile Ecosystems", 2008

[21] Online Services srl, “Constructive assembly and method for granting authorized access to an Internet service platform”, PCT/RO2017/000002 Patent pendig, 2016
[22] Entersekt, “OTP security past its expiration date” https://www.entersekt.com, 2014
[23] Entersekt, “Securing the mobile banking channel” https://www.entersekt.com, 2014
[24] Sonia Chiasson, “Usable Authentication and Click-Based Graphical Passwords” ISBN: 978-0-494-47475-4, 2009
[25] M. Mathuri Pandi, A. Valarmathi, “A secured graphical password authentication system” International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5 ISSN: 2278-0181, 2013
[26] Amish Shah, Partth Ved, Avani Deora, Arjun Jaiswal, Mitchell D'silva, "Shoulder-surfing resistant graphical password system" Procedia Computer Science 45, pp. 477 – 484, 2015
[27] Vinit Khetani, Jennifer Nicholas, Anuja Bongirwar, Abhay Yeole, “Securing web accounts using graphical password authentication through watermarking” International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 6 ISSN: 2231-2803, 2004
[28] "Logging in W3C httpd". World Wide Web Consortium. 1995-10-12. Retrieved 2018-11-20.
[29] R. Fielding et. al., “Hypertext Transfer Protocol – HTTP/1.1”, The Internet Society (1999).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 92-96 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

The Visualization of Cattle Movement Data in The State of Pará in 2016 Through Networks of Animal Transit Graphs and Guides

Samuel Carvalho de Aragão[1,*], Agnaldo Reis Pontes[1], Luis Manuel Borges Gouveia[2], Samuel Franco Lopes[3], Pier Kenji Rauschkolb Katsuda[4], Anirene Galvão Tavares Pereira[4], Márcio Teixeira Oliveira[5], Jefferson Pinto de Oliveira[6], Rita do Socorro Brito Coroa[6], Gilson Ferreira Araújo[6], Marcio Merêncio Panza de Siqueira[6]

[1]*Federal Institute of Education, Science and Technology of Pará – IFPA, 68627-100, Brazil.*

[2]*Fernando Pessoa University – UFP, 04249-004, Portugal.*

[3]*Faculty of Amazônia – UNAMA – 68010-200, Brazil.*

[4]*São Paulo State University (UNESP)- FMVA – 16050-680. Brazil.*

[5] *Federal Institute of Education, Science and Technology of Mato Grosso do Sul – IFMS, 79641-162. Brazil.*

[6] *Pará State Agricultural Defense Agency – ADEPARÁ, 66613-115, Brazil.*

ARTICLE INFO

Article history:
Received: 16 August, 2018
Accepted: 14 September, 2018
Online: 18 September, 2018

Keywords:
Network Theory
SNA,
Veterinary Epidemiology
Commercial Agglomerates

ABSTRACT

Animal movement is inherent in the marketing between the rural productive units, establishing space-time connections between them. The relational nature of such information is kept in the Animal Transit Guides (GTA), a mandatory issuance in Brazil. When evaluating such set of information, this work aimed at characterizing the bovine movement network in the state of Pará through the application of concepts from Social Network and Network Theory analysis. Where the nodes and edges that structure the network are respectively the rural properties and the number of cattle moved between them. The results obtained characterize the bovine movement in the state, which show distinct patterns such as the movement of a great number of animals rambling and a clear seasonality due to commercialization purposes. The evaluation of the network components' modularity also shows the subdivision of the state in four commercial clusters, where the central measures of its components define municipalities of economic and epidemiological interest. The results described herein help in the support of health surveillance and control measures, as well as the adoption of economic strategies and policies.

1. Introduction

Social Network Analysis (SNA) and the Theory of Networks are methodological concepts that describe the interactions between individuals in a group, as well as the collective behavior of a group [1]. Although widely adopted in the social sciences, psychology, and anthropology, SNA and the Theory of Networks have only recently been added to the analytical arsenal used to evaluate the nature and extent of relationships in the rural productive chain [2]. These methodologies allow researchers to evaluate the structure of the flow of animals between different productive units, and the impact that this has on the spread of a wide range of infectious agents [3–6]. SNA and the Theory of Networks also allow the extraction of characteristics of particular economic interest to the livestock industry, such as commercial patterns and commercial agglomerations [7–9].

The study of networks initially describes how the entities that teh networks are composed of are connected to each other, where the unit of interest is called the node and the connections between these are called edges [10]. Unlike in the human sciences, where the unit is the individual, in veterinary medicine the focus is on the collective unit. Municipalities and rural properties are represented by nodes and the number of animals moved between these are represented by the edges [11-12]. From an economic point of view, animal movement between productive units represents an

* Samuel Carvalho de Aragão, Email: samuel.aragao@ifpa.edu.br

important aspect of livestock commercialization, as it establishes extensive spatial connections between different units [9]. This movement also favors the occurrence and spread of diseases, as movement of animals is of central importance to the dynamics of diseases and occurrence of outbreaks [13,14]. In most industrialized countries, measures are adopted to identify, record, and trace animal movements as a prerequisite for the effective adoption of disease surveillance and control measures [15].

In Brazil, the Animal Transit Guide (GTA), instituted by Ordinance 22 on January 13th, 1995 [16] and updated in 2012 by Normative Instruction No. 18, dated July 18th, 2006 [17], is the official document mandatory for the intra and interstate movement of animals, regardless of the purpose of the movement. It is an important tool for sanitary control, and provides support for the prevention of diseases that affect domestic species of production [18]. Given the relational nature of the information contained in the transit guides, these are readily converted into the attributes that make up a network, where nodes represent the properties of origin and destination, and their edges represent the animal movement that occurs between them [19].

In the state of Pará, the issuance, control, and analysis of the GTAs, as well as adoption of the necessary measures to maintain the health of the livestock are the responsibility of the Agency of Agricultural Defense of the State of Pará (ADEPARÁ). ADEPARÁ is the government body responsible for animal health in the state of Pará. The agency monitors the fifth largest national cattle herd in Brazil, which comprises more than 19 million heads [20]. This monitoring is done according to norms established by the Ministry of Agriculture, Livestock and Supply (MAPA) [21].

The State of Pará, which contains approximately 9% of cattle in Brazil, and experienced a 43% increase in the number of cattle between 2004 and 2013, performed better than the states that include the largest national herds, such as Mato Grosso, Minas Gerais, Goiás and Mato Grosso do Sul [22]. Given the importance of cattle breeding not only at the state level, but at the national level, it is of prime importance to characterize animal movements among the structural components of the production chain. Such information will enable the implementation of effective measures for sanitary, market, and managerial control [23]. Based on the information described above, this work describes the main characteristics of cattle movement in the state of Pará. We applied network analysis techniques to the relational, temporal, and spatial components of the dataset formed by the GTAs issued during the year 2016 in the state of Pará.

2. Materials and Methods

2.1 Study area

The state of Pará is located in the Northern Region of Brazil, and covers an area of more than 1,247,000 km^2. The state of Pará has estimated population of over 8 million inhabitants, and in 2016 included more than 110,000 rural properties [24].

2.2 Animal Transit Guides

The information used to compose the cattle handling network was made available by ADEPARÁ, and includes the compilation of all GTAs issued during the year 2016 in the state of Pará. Each GTA issued has a unique identifier associated with the date, party, and issuing regional unit. Information relevant to the characterization of the movement carried out includes the type of transport, the species transported, the purpose of the transport, the number of animals transported according to sex and age group (only for cattle and buffalo), municipalities and federal units of origin and destination, as well as information identifying the properties or establishments from where and to which the animals were moved. The information comprised by the CPF/CNPJ (Physical Person Registration/National Registration 0f Legal Entities), name of the owner, and name of the property or establishment of origin or destination, is assigned a unique identifier established by ADEPARÁ. Spatial information, i.e. the latitude and longitude of the municipalities of origin and destination were added to the dataset.

2.3 Cattle Handling Network

The network structure is derived from the relational information contained in each GTA issued, where the nodes represent the productive units, and the edges represent the cattle movement between them. The identification of each node corresponds to that adopted by ADEPARÁ and its attributes include the name, federative unit, and geographical position of the municipality to which it belongs. The same strategy was adopted to identify the edges, keeping the unique identifier designated by ADEPARÁ to each GTA issued. The attributes include the date of issue, the temporal component of the network, the purpose and type of transport, the number of cattle transported according to age and sex, and the identification of the origin and destination nodes. The movement described by each edge carries implicit direction information, starting at the origin and ending at the destination. Therefore, node A connected to node B, represented by A-B, is different from the connection B-A, where the distance between them is defined as the lowest number of directed edges necessary to connect them.

The network properties and characteristics were evaluated, including their diameter, modularity, heavy grades of input and output, centrality of intermediation, proximity centrality, PageRank, and clustering coefficient. The network diameter is defined as the path with the greatest number of edges between two nodes A and B of the network [25]. Modularity is a measure of the quality of the clustering of a given set of nodes [26]. The number of edges arriving at a given node and the number of edges that start from this are respectively the degrees of entry and exit [25]. The weight or multiplier of these degrees was defined as the number of animals moved between two nodes [27].

The number of shortest paths among all nodes of the network passing through node A is defined as intermediation centrality of node A [28]. The inverse of the sum of all distances from node A to all other nodes of the network defines the centrality of its proximity [29]. The PageRank of node A corresponds to the probability of arriving at this node starting from a random network node B [30]. The clustering coefficient indicates the degree to which the network nodes tend to cluster [31].

The descriptive analysis, as well as conversion of the information made available by ADEPARÁ, in the set of nodes and edges that make up the network were performed using development codes in R statitistical software [32], with RStudio 2.11.1+ [33]. The analysis and extraction of the network characteristics described above were performed using the R package ‘iGraph’ [34] and the open access software Gephi 0.9.1 [35].

Table 1: Number and percentage of cattle and GTAs issued, according to the purpose of cattle movement

PURPOSES	Cattle (n)	Cattle (%)	GTA (n)	GTA (%)
SLAUGHTER	3515074	31.63	210438	48.54
FATTEN	6803942	61.23	193063	44.53
SPORTS	3267	0.03	139	0.03
EXPORT	174604	1.57	7733	1.78
EXHIBITION	2103	0.02	190	0.04
AUCTION	131214	1.18	5115	1.18
RESEARCH	1	~0	1	~0
Quarantine	243677	2.19	8105	1.87
Reproduction	182671	1.64	5819	1.34
Return of Agglomeration	55234	0.5	2939	0.68

3. Results and Discussion

With 135 regional management units and 1,841 users registered in the system, the State Agency of Agricultural and Livestock Defense of the State of Pará issued 433,545 GTAs for cattle between January 1st and December 31st, 2016. The animals transported had origins or destinations in one of the 82.010 registered properties, which were distributed in 929 municipalities of 24 federative units. In Pará, there are 78.362 properties distributed in 142 municipalities, with more than 32% of the state's properties concentrated in only 10 municipalities.

Over 90% of GTAs involved road transport as a means for the movement of cattle (Table 1). This number corresponds to the distribution of the number of animals transported by GTAs, with a median of 25 cattle, which corresponds to the maximum vehicle load capacity. Road transport is an important factor in determining strategies for epidemiological control, given the possibility of rapid spread of diseases by movements over large territories [9]. In spite of the number of animals transported by road, a significant number of cattle (over 3 million) were moved on foot (Table 2), with an approximate average of 126 animals per GTA issued. In this type of movement, the animals are moved in large numbers over short distances, remaining in a region close to the origin, a factor that contributes to the local spread of diseases [36].

Table 2: Number and percentage of cattle and GTAs according to the type of transport

Type of Transportation	Cattle (n)	Cattle (%)	GTA (n)	GTA (%)
ON FOOT	3299532	29.69	26138	6.03
RAILWAY	1059	0.01	32	0.01
MARITIME / FLUVIAL	329155	2.96	13648	3.15
ROAD	7482269	67.33	393727	90.82

In Pará, beef cattle farming is very important, as shown in Table 1, and 31% of cattle moved in the state are destined for slaughter, and 61% of cattle are moved for fattening. The low index of animals moved for breeding (1.64%) is related to two marketing moments, after rearing and after fattening. The number of cattle moved for quarantine purposes (2.19%) is related to the export of live cattle, with 1.57% of the cattle moved for export purposes (Table 1). This information is also supported by the temporal evaluation of the GTAs issued (Figure 1 and Figure 2), which indicates an increase in the number of cattle destined for quarantine during the months of May and June, which precedes the increase of GTAs issued for export during the months of June and July, which is end of the rainy season, and the period that cattle gain the most weight.

Figure 1 shows a clear seasonality of cattle breeding in Pará, with an increase in the number of cattle destined for fattening during the months of April and October, and a progressive increase in the number of cattle slaughtered between May and August. Cattle moved for fattening in October are slaughtered in early May, and those moved for fattening in April are slaughtered at the end of the rainy season in mid-August. Cattle moved for exhibitions and sports are concentrated in the months of July and August (Fig. 1 and Fig. 2), which are critical time points for the monitoring of these cattle in order to prevent the spread of diseases [2].

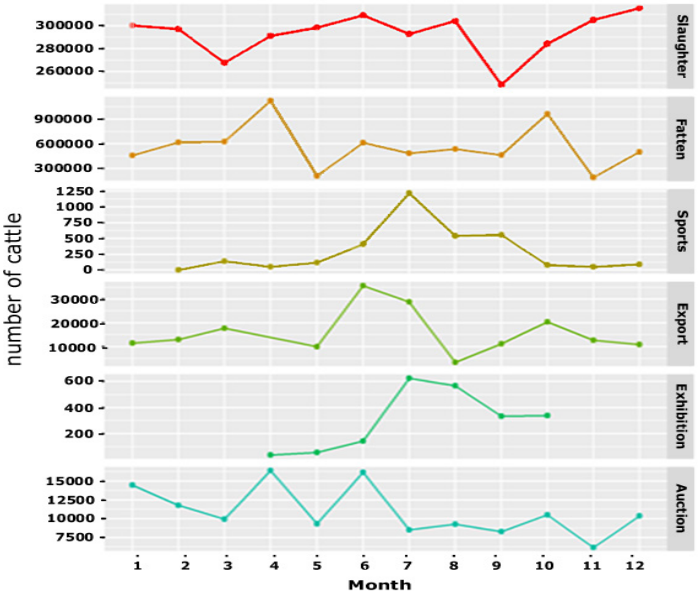


Figure 1 - Number of cattle moved during the year 2016 according to the purpose of the GTA.

Table 3:Number and percentage of cattle moved according to age and sex.

Age (months) and Sex	Bov (n)	Bov (%)
0 - 12 (males)	1496779	13.47
0 - 12 (females)	378254	3.40
13 - 24 (males)	1510723	13.60
13 - 24 (females)	705692	6.35
25 - 36 (males)	1428094	12.85
25 - 36 (females)	801877	7.22
> 36 (males)	1812238	16.31
> 36 (females)	2978358	26.80
Total (males)	6247834	56.23
Total (females)	4864181	43.77

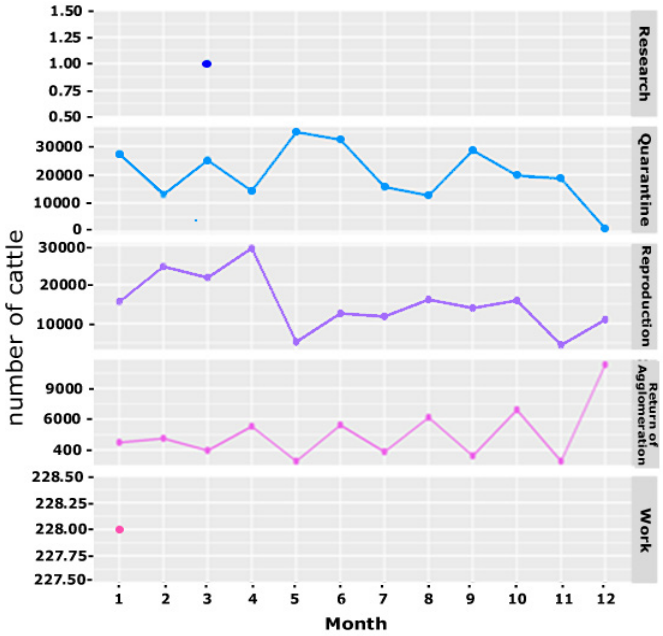


Figure 2 - Number of cattle moved during the year 2016 according to the purpose of the GTA.

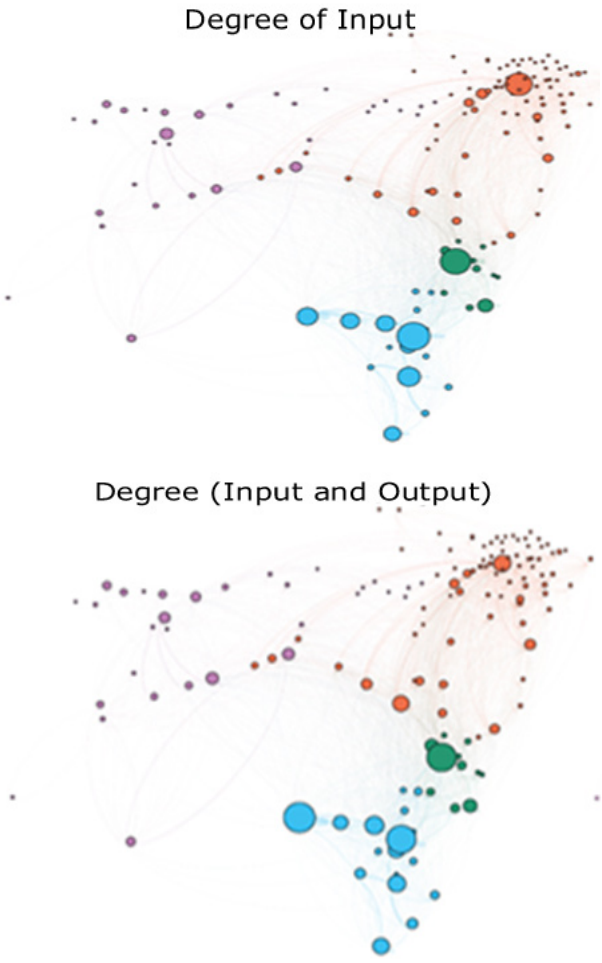


Figure 3 - Properties of the cattle movement network in the state of Pará according to the modularity class to which the municipality belongs.

Figure 4 - Properties of the cattle movement network in the state of Pará according to the modularity class to which the municipality belongs

The commercialization of male cattle in Pará is independent of age. The high number of females of reproductive age transported for slaughter (Table 3) and export purposes (Table 1), has a direct impact on the number of cattle in the state, reducing the population growth rate, and keeping the number of cattle stable. In disease surveillance programs, categories of epidemiological importance are those composed of animals susceptible to disease [37]. Therefore, determining the sample group of interest reduces the number of samples necessary for statistical confirmation of the establishment of disease in the population, as well as its spread, reducing the time required for detection and response by the responsible health agency [23].

The diameter of the net of cattle movement obtained is equal to 5, which, in other words, indirectly represents the number of municipalities through which an animal must pass to cover the largest geodesic distance of the net. Using the latitude and longitude information of each municipality, the average distance between the municipalities is approximately 220 km, with a magnitude of approximately 8.736 km, which is the sum of all distances. These results highlight the vast territory of the state.

4. Conclusion

Although the analyses and metrics evaluated were used in the state of Pará, they can be applied and expanded to other states, as well as to other production species. The quantitative description of the cattle movement network proposed in this study allowed us to reveal factors that shape the commercialization and displacement patterns of cattle in Pará. This data can contribute to the development of effective strategies to control and monitor the diseases that occur in cattle.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] B. Martínez-López, A. M. Perez, and J. M. Sánchez-Vizcaíno, "Social Network Analysis. Review of General Concepts and Use in Preventive Veterinary Medicine," *Transbound. Emerg. Dis.*, vol. 56, no. 4, pp. 109–120, May 2009.

[2] J. L. Hardstaff, B. Häsler, and J. R. Rushton, "Livestock trade networks for guiding animal health surveillance," *BMC Vet. Res.*, vol. 11, no. 1, p. 82, 2015.

[3] F. Natale, L. Savini, A. Giovannini, P. Calistri, L. Candeloro, and G. Fiore, "Evaluation of risk and vulnerability using a Disease Flow Centrality measure in dynamic cattle trade networks," *Prev. Vet. Med.*, vol. 98, no. 2, pp. 111–118, 2011.

[4] I. Z. Kiss, D. M. Green, and R. R. Kao, "The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread," *J. R. Soc. Interface*, vol. 3, no. 10, pp. 669–677, 2006.

[5] H. H. K. Lentz *et al.*, "Disease Spread through Animal Movements: A Static and Temporal Network Analysis of Pig Trade in Germany," *PLoS One*, vol. 11, no. 5, pp. 1–32, 2016.

[6] B. Shirouchi *et al.*, "Fatty acid profiles and adipogenic gene expression of various fat depots in Japanese Black and Holstein steers," *Meat Sci.*, vol. 96, no. 1, pp. 157–164, 2014.

[7] E. E. Gorsich *et al.*, "Mapping U.S. cattle shipment networks: Spatial and temporal patterns of trade communities from 2009 to 2011," *Prev. Vet. Med.*, vol. 134, no. Supplement C, pp. 82–91, 2016.

[8] M. C. Vernon and M. J. Keeling, "Representing the UK{\textquoteright}s cattle herd as static and dynamic networks," *Proc. R. Soc. London B Biol. Sci.*, vol. 276, no. 1656, pp. 469–476, 2009.

[9] P. Bajardi, A. Barrat, F. Natale, L. Savini, and V. Colizza, "Dynamical Patterns of Cattle Trade Movements," *PLoS One*, vol. 6, no. 5, pp. 1–19, 2011.

[10] C. Dubé, C. Ribble, D. Kelton, and B. McNab, "A Review of Network Analysis Terminology and its Application to Foot-and-Mouth Disease Modelling and Policy Development," *Transbound. Emerg. Dis.*, vol. 56, no. 3, pp. 73–85, Apr. 2009.

[11] S. Schärrer *et al.*, "Evaluation of farm-level parameters derived from animal movements for use in risk-based surveillance programmes of cattle in Switzerland," *BMC Vet. Res.*, vol. 11, no. 1, p. 149, Jul. 2015.

[12] F. Kschischang, "An Introduction to Network Coding," 2012.

[13] S. E. Robinson, M. G. Everett, and R. M. Christley, "Recent network evolution increases the potential for large epidemics in the British cattle population," *Journal of the Royal Society Interface*, vol. 4, no. 15. London, pp. 669–674, Aug-2007.

[14] A. Ortiz-Pelaez, D. U. Pfeiffer, R. J. Soares-Magalhães, and F. J. Guitian, "Use of social network analysis to characterize the pattern of animal movements in the initial phases of the 2001 foot and mouth disease (FMD) epidemic in the UK," *Prev. Vet. Med.*, vol. 76, no. 1, pp. 40–55, 2006.

[15] P. Motta *et al.*, "Implications of the cattle trade network in Cameroon for regional disease prevention and control," *Sci. Rep.*, vol. 7, p. 43932, Mar. 2017.

[16] Brasil, "Portaria no 22, de 13 de janeiro de 1995," 1995.

[17] Brasil, "Decreto n° 5.741, de 03 de março de 2006," 2006.

[18] OIE OM de SA, "Código sanitário para os animais terrestres," 2015.

[19] M. Amaku *et al.*, "Infectious disease surveillance in animal movement networks: An approach based on the friendship paradox," *Prev. Vet. Med.*, vol. 121, no. 3, pp. 306–313, 2015.

[20] ADEPARA, "Agência Estadual de Defesa Agropecuária do Estado do Pará," 2015.

[21] IBGE, "Instituto Brasileiro de Geografia e Estatística."

[22] FAPESPA, "BOLETIM AGROPECUÁRIO DO ESTADO DO PARÁ," 2015.

[23] K. L. VanderWaal *et al.*, "Network analysis of cattle movements in Uruguay: Quantifying heterogeneity for risk-based disease surveillance and control," *Prev. Vet. Med.*, vol. 123, no. Supplement C, pp. 12–22, 2016.

[24] IBGE, "Sistema IBGE de Recuperacao automatica–SIDRA." 2015.

[25] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Soc. Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[26] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, 2006.

[27] R. R. Kao, L. Danon, D. M. Green, and I. Z. Kiss, "Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain," *Proc. R. Soc. London B Biol. Sci.*, vol. 273, no. 1597, pp. 1999–2007, 2006.

[28] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[29] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[30] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Natl. Acad. Sci.*, vol. 105, no. 4, pp. 1118–1123, 2008.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, p. 440, Jun. 1998.

[32] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *J. Comput. Graph. Stat.*, vol. 5, no. 3, pp. 299–314, 1996.

[33] J. S. Racine, "RStudio: A Platform-Independent IDE for R and Sweave," *J. Appl. Econom.*, vol. 27, no. 1, pp. 167–172, 2012.

[34] G. Csardi and T. Nepusz, "The igraph software package for complex network research," 2006.

[35] M. Bastian, S. Heymann, and M. Jacomy, "Gephi : An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs," 2017.

[36] C. Ensoy, C. Faes, S. Welby, Y. Van der Stede, and M. Aerts, "Exploring cattle movements in Belgium," *Prev. Vet. Med.*, vol. 116, no. 1, pp. 89–101, 2014.

[37] L. J. Hoinville *et al.*, "Proposed terms and concepts for describing and evaluating animal-health surveillance systems," *Prev. Vet. Med.*, vol. 112, no. 1, pp. 1–12, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 101-113 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimal Designs of Constrained Accelerated Life Testing Experiments for Proportional Hazards Models

Xiaojian Xu[*,1], Wanyi Huang[2]

[1]*Department of Mathematics and Statistics, Brock University, L2S 3A1, Canada*

[2]*XPO Logistics, Changhua 500, Taiwan*

A R T I C L E I N F O

Article history:
Received: 30 November, 2018
Accepted: 06 January, 2019
Online : 29 January, 2019

Keywords:
Multiple step-stress proportional hazards model
Fisher's information
Optimal ALT design
Design constraint Asymptotic covariance matrix
Exact design
D-optimal design
A-optimal design
Q-optimal design

A B S T R A C T

This paper investigates the methods of optimal design construction for step-stress accelerated life testing (ALT) when a Cox's hazards model is adopted with either a linear or a quadratic baseline hazard function. We discuss multiple step-stress plans for time-censored ALT experiments. The maximum likelihood method is utilized for estimating the model parameters. The information matrices have been derived for both models. The optimal stress-changing times and optimal stress levels are determined simultaneously under three different optimality criteria. In order to demonstrate the performance of the resulting designs, a simulation procedure is also provided. The efficiencies of our resulting optimal three-step-stress ALT plans are compared with their competitors using both practical examples and a simulation study. The efficiency comparison results have shown that the three-step-stress designs obtained with two optimal stress changing times and an optimal middle stress level are most efficient, compared to the corresponding optimal two-step-stress designs and to the optimal three-step-stress designs with a conveniently chosen middle stress. Furthermore, the efficient gains are most significant for hazard rate prediction for both cases when either a linear or a quadratic baseline hazard is assumed. Additionally, such efficiency gain is much greater for the case when the baseline function being quadratic than the case when that being simple linear.

1. Introduction

Reliability significantly influences the quality of product. Thus, many manufacturers make great effort to enhance the product reliability that largely determines the product competitiveness. Such importance has brought practitioners' attention to reliability evaluation. In life data analysis, we have to collect the failure time of a product under normal design conditions to quantify life characteristics of the product. However, such failure data for lifetime are very difficult to obtain in many situations, especially for the product with high reliability. Nowadays, lifetimes of many products are too long and the life testing period between design and release is limited; so, the tests under normal design conditions are too lengthy to get any failures. To overcome this problem, accelerate life testing (ALT) has been developed.

Since ALT can shorten the lifetime of a product, we often adopt it in order to obtain the failure information quickly within a limited time frame. In an ALT experiment, the test units are generally subjected to the stress levels that are higher than normal design level. The commonly used stress factors for failure acceleration include temperature, vibration, voltage, and pressure. Both the use of specific accelerated stress factors and the range of stress levels for a particular material or product are often suggested by engineering practice. Then, the failure data obtained at accelerated conditions have to be extrapolated through a proper model so that the characteristics of life distribution at normal design conditions can be estimated. A number of different types of stress loading schemes (such as constant, cyclic, step, progressive, and random stress loading) are available in practice when performing an ALT. By the relation between the stress levels and testing time, these stress loading schemes can be classified into two categories: time-independent and time-dependent stress loadings. When the stress loading is time-independent such as constant stress loading, the

[*]Corresponding Author: Xiaojian Xu, 1812 Sir Isaac Brock Way, St. Catharines, ON, L2S 3A1 Canada; 905-688-5550 ext. 3300; xxu@brocku.ca.

stress level applied to each of the test units stays the same during the entire testing period. One of the demerits of using constant stress loading in ALT is that it could still take too long to run for the test experiment to observe sufficient failures when the inappropriate testing stress levels are applied. To address this problem, a time-dependent stress loading scheme is preferred to assure quick failures. When a time-dependent stress loading, such as step-stress, is adopted, test units are subjected to a stress level that is changing over testing time. In this paper, we consider step-stress ALT where all test units are subjected to the stress levels increasing by steps.

This paper is an extension of our previous work [1] originally presented in The Third IEEE International Conference on World Computing and Big Data Analysis. In literature, almost all the previous works done by others for Cox's proportional hazards (PH) based ALT models are very limited on simple step-stress assuming the baseline hazard function is simple linear. However, [2] have indicated that the optimal multiple step-stress ALT may further improve the quality of the reliability pre diction for a parametric model. Therefore, this paper broadens the results of [1] where has only addressed the optimal ALT plans for PH models with a linear baseline function and moves on to discussing multiple step-stress optimal designs for ALT when adopting a PH model with either a simple linear or a quadratic baseline hazard function. In this paper, both optimal stress levels and optimal stress-changing times for three-step-stress ALT designs are derived under three different optimality criteria.

2. Literature Review and Preliminaries

2.1. Optimal designs for step-stress ALTs with a non-PH model

Miller and Nelson [3] first discussed Q-optimal designs for simple step-stress ALT tests with complete failure data assumed to be exponentially distributed. Their optimal designs were attained by minimizing the asymptotic variance (AVAR) of the maximum likelihood estimator (MLE) of the mean lifetime at the normal design stress level. Then, their work was extended to censored data by [4] who obtained the optimal simple step-stress ALT designs incorporating time-censoring. For some products or material, their failure times often follows a Weibull distribution. Assuming a Weibull distribution with a constant scale parameter, both [5] and [6] constructed optimal simple step-stress ALT designs for time-censoring. Bai and Kim in [5] obtained the optimal low stress level and optimal stress-changing time in order to minimize the AVAR of the MLE of a specific quantile of the product's lifetime distribution at the normal design stress level whereas [6] obtained their optimal stress-changing time in order to minimize the AVAR of the MLE for reliability prediction instead. In addition, Hunt and Xu in [7] further investigated optimal simple step-stress ALT plans for a Weibull distribution; however, they assumed both the shape and scale parameters were functions of the stress levels. Their resulting optimal designs chosen the stress-changing time in order to minimize AVAR of the MLE of reliability prediction at the normal design stress level and at a pre-specified time. They also reviewed the research work on optimal designs for step-stress ALT. Please see the references therein. We note that all these work previously done had provided the design construction methods only for simple step-stress ALT plans.

Moreover, Ma and Meeker in [8] extended the research by [5] to provide a general method for multiple step-stress ALTs assuming a log-location-scale family of distributions. They discussed an approach to calculate the large-sample approximate variance of the MLE for a percentile of the failure time distribution at normal design conditions when the failure data were observed from a step-stress ALT. By adopting a cumulative exposure model, their approach allowed for both multiple step-stress loading and censoring. Their results also showed that depending on the values of the model parameters and certain percentile of interest, one of the three test plans proposed could be the most preferable in terms of optimum variance. For a Weibull lifetime distribution, however, with possible inaccuracy in the assumed log-linear life-stress [2] investigated the optimal stress-changing time for simple step-stress ALT plans in order that the asymptotic mean squared error of the underlying reliability estimator could be minimized, and the robust choices of three-step-stress plans were also discussed with the awareness of possible imprecision in the assumed life-stress relationship by minimizing the asymptotic squared bias.

2.2. Optimal designs for constant stress or simple step-stress ALTs with a PH model

Jiao in [9] first investigated the optimal design problem for a PH model when a constant-stress ALT experiment being planned, and then developed the optimal designs for reliability prediction by optimally choosing both stress levels and proportion of units allocated to each stress level in order to attain the most accurate reliability estimate at normal design conditions. Moreover, when a step-stress ALT experiment was planned, [9] discussed the simple step-stress ALT plan for reliability prediction and obtained the optimal stress level by minimizing the variance of the MLE of hazard rate at the normal design stress level and over a pre-specified time period. In addition, [9] also provided an algorithm for solving the constrained nonlinear optimization problems.

On the other hand, Elsayed and Zhang in [10] revealed an optimal simple step-stress ALT plan so as to obtain the most accurate reliability function estimates at normal design conditions. They also formulated a nonlinear programming problem to minimize the asymptotic variance of the hazard rate estimator over a prespecified the period at the normal design stress level. More recently, Hu, Plante, and Tang in [11] briefly discussed the optimal low and high stress levels in a simple step-stress ALT in order to minimize the mean squared error of the estimated upper confidence bound for the cumulative failure probability of a product at normal design conditions, with a given stress-changing time. In sum, all these existing works provided the methods of optimal design construction only for simple step-stress ALT plans. Therefore, we expand the previous work of others and investigate the optimal designs of multiple step-stress ALT for PH models in this paper.

2.3. Optimal designs for general proportional hazards models

There is broader literature of optimal designs available for general proportional hazards models, which are not necessary with consideration of step-stress ALT plans. To name a few, Becker, McDonald, and Khoo in [12] constructed D-optimal designs for proportional hazards models with one or two parameters when its baseline hazard function was specified. They developed the

method of minimum variance design construction when various censoring schemes were adopted. Dette and Sahm in [13] provided a standardized maximum variance design criterion which could be applied to obtain the optimal designs whereas McGree and Eccleston in [14] created compound criteria so that optimal designs could be derived for multi-objective scenarios. L´opez-Fidalgo, Rivas-L´opez, and Del Campo in [15] also proposed an algorithm to find optimal designs for typical Cox regression models incorporating censoring. More recently, Konstantinou, Biedermann, and Kimber investigated the general maximin D-optimal designs for a class of models and discussed the application of their design construction method to the proportional hazards models in [16].

2.4. Our model and some preliminaries

One of the most commonly used means for predicting the lifetime of a product is Cox's PH model since it provides sufficient flexibility for identifying the effects of covariates on the failure rate. In this paper, a PH model with hazard ratio being independent of time is considered. Therefore, the hazard rate of a product can be expressed

$$\lambda(t;\mathbf{s}) = \lambda_0(t)\exp\left(\boldsymbol{\beta}^T\mathbf{s}\right), \tag{1}$$

where $\lambda_0(t)$ is a baseline hazard function, $\boldsymbol{\beta}$ is a column vector of unknown parameters, and $\mathbf{s}$ is a column vector of the covariates (applied or transformed stresses) for an ALT experiment, and that are independent of the baseline hazard. Presumably, the stresses are having multiplicatively effects on the hazard rate in this model.

We conduct ALT with step-stress loading where all the test units begin at a prespecified stress level. After a time fraction, the stress level is changed to a higher stress level. The stress level can be raised more than once before the test completes. A simple step-stress ALT, also called a two-step-stress ALT, only uses two stress levels during an ALT experiment. In contrast, in multiple step-stress ALT, more than two stress levels are needed and the stress level is raised at least twice before the test ends. In this paper, we discuss multiple step-stress ALT experiments. We denote by s_D the normal design stress level. For a three-step-stress, we signify the low, middle and high stress levels by s_1, s_2, and s_3, the stress-changing times by τ_1 and τ_2, and censoring time by c.

In this paper, we assume $\lambda_0(t)$ to be in either a linear or a quadratic form. For either case, we investigate the optimal three-step-stress ALT design construction under D-optimality, A-optimality, and Q-optimality, respectively. We obtain D-optimal designs in order to minimize the determinant of covariance matrix of the estimators for the model parameters (or equivalently maximize the determinant of Fisher's information matrix), A-optimal designs in order to minimize the trace of covariance matrix of the estimators for the model parameters, and Q-optimal design to minimize the asymptotic variance of the estimator for a specific quantity of interest. The major quantity of interest in this paper is the average hazard rate over a particular period of time under normal design conditions.

For the case of $\lambda_0(t)$ being simple linear, the main part of the optimal design construction was presented in [1]. Therefore, we will only restate the necessary notations being continuously used here and also summarize the previous results for comparison purpose in this paper. For the case of $\lambda_0(t)$ being a quadratic function, we will present the discussion and derivation in full. Furthermore, because there was no simulation study done previously for either case, we will demonstrate simulation and comparison studies for both cases in the present paper.

The rest of this paper is organized as follows: in Section 3, the optimal three-step-stress ALT design results have been summarized for the case when the baseline hazard function $\lambda_0(t)$ in (1) is assumed to be a simple linear function. In Section 4, the optimal three-step-stress ALT design have been derived for the case when $\lambda_0(t)$ is considered to be a quadratic function. The method of the optimal design construction involves the minimization for a nonlinear objective function with nonlinear constraints, and some practical examples are used to illustrate the proposed method for the construction of constrained optimal designs in Sections 3 and 4. We have also evaluated the performance of the resulting designs obtained in both Sections 3 and 4 through simulations and comparisons in Section 5. Some concluding remarks are presented in Section 6.

3. Optimal designs when $\lambda_0(t)$ being a simple linear function

Xu and Huang in [1] first focused on determining the optimal three-step-stress ALT designs for a PH model with time-censoring in the case when the baseline hazard function is considered being a simple linear function. They have provided detailed derivation for the information matrix and the optimal choices of both the middle stress level and stress-changing times for all cases considered there. For comparison reasons, we keep the same notations which are presented in Subsection 3.1, and their main results are summarized in this section.

3.1. Notation

Although the method of development for multiple step-stress ALT designs can be provided in general, we formulate our design construction using three-step-stress ALT experiments for its simplicity. A three-step-stress ALT experiment with time-censoring involves a predefined censoring time c, and three test stress levels s_1, s_2, and s_3, which satisfies $s_D < s_1 < s_2 < s_3$. Assuming there are n test units available for the ALT experiment, they are all first placed at low stress level s_1 for a time interval $[0, \tau_1]$. Afterwards, the test units survived by time τ_1 are subjected to a middle stress level s_2 for the next time interval $(\tau_1, \tau_2]$. Next, the remaining units by time τ_2 are subjected to the highest stress level s_3 for the last time interval $(\tau_2, c]$. Then, the test ends at censoring time c. We consider Model (1), under a stress level s, where $\lambda_0(t)$ being independent of s.

We make use of the notations $F(t;s)$, $f(t;s)$, $\Lambda(t;s)$ $R(t;s)$

t, respectively, at a given stress level s. We also employ the cumulative exposure model (CEM), please see [17] for details, to address the changes

of $F(t;s)$, in lifetime t due to the stress step-ups in a step-stress ALT experiment. We denote the cdf under s_i by

$$F_i(t)=F(t\,|\,s=s_i),\ i=1,2,\text{ and }3.$$

Then, the stepwise cdf with respect to t can be expressed by:

$$F_{CEM}(t)=\begin{cases} F_1(t) & \text{if } t\le\tau_1,\\ F_2(a+t-\tau_1) & \text{if } \tau_1\le t\le\tau_2,\\ F_3(b+t-\tau_2) & \text{if } \tau_2\le t\le c,\\ P(t>c) & \text{if } t>c,\end{cases}$$

Where a and b satisfy $F_1(\tau_1)=F_2(a)$ and $F_2(a+\tau_2-\tau_1)=F_3(b)$. Namely, $a=F_2^{-1}\left[F_1(\tau_1)\right]$ and $b=F_3^{-1}\left[F_2(a+\tau_2-\tau_1)\right]$.

We also make use of the following three indicators, each as a function of stress change times τ_1, τ_2, or censoring time c, and failure time t :

$$I_1=I_1(t\le\tau_1)=\begin{cases}1 & \text{if } t\le\tau_1,\\ 0 & \text{if } t>\tau_1,\end{cases}\quad I_2=I_2(t\le\tau_2)=\begin{cases}1 & \text{if } t\le\tau_2,\\ 0 & \text{if } t>\tau_2,\end{cases}$$

$$I_3=I_3(t\le c)=\begin{cases}1 & \text{if } t\le c,\\ 0 & \text{if } t>c,\end{cases}$$

where $0<\tau_1<\tau_2<c$. Taking a three-step-stress design ξ with three stress levels being s_1, s_2, and s_3, the log-likelihood function of an observed lifetime t can be written as:

$$\begin{aligned}\ln L(t;\xi)=\ & I_1I_2I_3\left[\ln(\gamma_0+\gamma_1t)+\beta s_1-\left(\gamma_0t+\frac{\gamma_1}{2}t^2\right)\exp(\beta s_1)\right]\\ &+(1-I_1)I_2I_3\left[\ln(\gamma_0+\gamma_1x)+\beta s_2-\left(\gamma_0x+\frac{\gamma_1}{2}x^2\right)\exp(\beta s_2)\right]\\ &+(1-I_2)I_3\left[\ln(\gamma_0+\gamma_1y)+\beta s_3-\left(\gamma_0y+\frac{\gamma_1}{2}y^2\right)\exp(\beta s_3)\right]\\ &-(1-I_3)\left[\left(\gamma_0d+\frac{\gamma_1}{2}d^2\right)\exp(\beta s_3)\right],\end{aligned}$$

where $x=a+t-\tau_1$, $y=b+t-\tau_2$, and $d=b+c-\tau_2$. (2)

As assumed the baseline hazard rate being independent of the covariates, according to [10], the Fisher's information matrix, F, for the full sample with size n can be obtained as

$$\mathbf{F}=n\begin{bmatrix} E\left\{-\frac{\partial^2\ln L}{\partial\gamma_0^2}\right\} & E\left\{-\frac{\partial^2\ln L}{\partial\gamma_0\gamma_1}\right\} & 0\\ E\left\{-\frac{\partial^2\ln L}{\partial\gamma_0\gamma_1}\right\} & E\left\{-\frac{\partial^2\ln L}{\partial\gamma_1^2}\right\} & 0\\ 0 & 0 & E\left\{-\frac{\partial^2\ln L}{\partial\beta^2}\right\}\end{bmatrix},$$

and the covariance matrix, Σ, of the maximum likelihood estimators (MLE) of γ_0, γ_1, β is the inverse matrix of $\mathbf{F}$:

$$\mathbf{\Sigma}=\mathbf{F}^{-1}=\begin{bmatrix} Var(\hat{\gamma}_0) & Cov(\hat{\gamma}_0,\hat{\gamma}_1) & 0\\ Cov(\hat{\gamma}_0,\hat{\gamma}_1) & Var(\hat{\gamma}_1) & 0\\ 0 & 0 & Var(\hat{\beta})\end{bmatrix}.$$

The elements of the Fisher's information matrix for t are the negative expectations of the corresponding second partial derivatives. We denote the non-zero elements of Fisher's information by F_{00}, F_{11}, F_{01}, and F_β. In [1], Xu and Huang have derived the expressions of these elements as follows:

$$\begin{aligned}F_{00}=&E\left\{-\frac{\partial^2\ln L}{\partial\gamma_0^2}\right\}\\ =&\exp(\beta s_1)\int_0^{\tau_1}\frac{1}{\lambda_0(t)}\exp\left[-\Lambda(t;s_1)\right]dt\\ &+\exp(\beta s_2)\int_a^{a+\tau_2-\tau_1}\frac{1}{\lambda_0(x)}\exp\left[-\Lambda(x;s_2)\right]dx\\ &+\exp(\beta s_3)\int_b^{b+c-\tau_2}\frac{1}{\lambda_0(y)}\exp\left[-\Lambda(y;s_3)\right]dy,\end{aligned}\tag{3}$$

$$\begin{aligned}F_{11}=&E\left\{-\frac{\partial^2\ln L}{\partial\gamma_1^2}\right\}\\ =&\exp(\beta s_1)\int_0^{\tau_1}\frac{t^2}{\lambda_0(t)}\exp\left[-\Lambda(t;s_1)\right]dt\\ &+\exp(\beta s_2)\int_a^{a+\tau_2-\tau_1}\frac{x^2}{\lambda_0(x)}\exp\left[-\Lambda(x;s_2)\right]dx\\ &+\exp(\beta s_3)\int_b^{b+c-\tau_2}\frac{y^2}{\lambda_0(y)}\exp\left[-\Lambda(y;s_3)\right]dy,\end{aligned}\tag{4}$$

$$\begin{aligned}F_{01}=&E\left\{-\frac{\partial^2\ln L}{\partial\gamma_0\gamma_1}\right\}\\ =&\exp(\beta s_1)\int_0^{\tau_1}\frac{t}{\lambda_0(t)}\exp\left[-\Lambda(t;s_1)\right]dt\\ &+\exp(\beta s_2)\int_a^{a+\tau_2-\tau_1}\frac{x}{\lambda_0(x)}\exp\left[-\Lambda(x;s_2)\right]dx\\ &+\exp(\beta s_3)\int_b^{b+c-\tau_2}\frac{t}{\lambda_0(y)}\exp\left[-\Lambda(y;s_3)\right]dy,\end{aligned}\tag{5}$$

and

$$\begin{aligned}F_\beta=&E\left\{-\frac{\partial^2\ln L}{\partial\beta^2}\right\}=s_1^2\exp(\beta s_1)\int_0^{\tau_1}\lambda_0(t)\Lambda(t;s_1)\exp\left[-\Lambda(t;s_1)\right]dt\\ &+s_2^2\exp(\beta s_2)\int_a^{a+\tau_2-\tau_1}\lambda_0(x)\Lambda(x;s_2)\exp\left[-\Lambda(x;s_2)\right]dx\\ &+s_3^2\exp(\beta s_3)\int_b^{b+c-\tau_2}\lambda_0(y)\Lambda(y;s_3)\exp\left[-\Lambda(y;s_3)\right]dy+s_3^2\Lambda(d;s_3).\end{aligned}\tag{6}$$

We note that the derivation of these elements in this subsection and the expressions of the loss functions in next subsection have been omitted, please see [1] for details.

3.2. Loss functions and design constraints

In this paper, we consider the following three different design criteria: D-optimality, A-optimality, and Q-optimality. In [1], Xu and Huang has derived the corresponding three loss functions as

$$L_D = \frac{1}{(F_{00}F_{11} - F_{01}^2)F_\beta}, \tag{7}$$

$$L_A = \frac{F_{00} + F_{11}}{F_{00}F_{11} - F_{01}^2} + \frac{1}{F_\beta}, \tag{8}$$

and

$$L_Q = \frac{F_{11}T - F_{01}T^2 + \frac{1}{3}F_{00}T^3}{F_{00}F_{11} - F_{01}^2} + \frac{s_D^2\left(\gamma_0^2 T + \gamma_0\gamma_1 T^2 + \frac{1}{3}\gamma_1^2 T^3\right)}{F_\beta}, \tag{9}$$

respectively.

In ALT practice, there are often some requirement needed on the minimum number of failures at each stress level. Similarly to [10] for designing optimal simple step-stress ALTs, we take certain practical constraints into consideration in our optimal design construction process for multiple step-stress ALTs. In this paper, we consider three design constraints as below:

- The minimum expected number of failures (MNF) at stress level s_1 is required as w_1:

$$n\Pr\left[t \le \tau_1 \middle| s_1\right] \ge w_1; \tag{10}$$

- The MNF at stress level s_2 is given as w_2 :

$$(n - n_1)\Pr\left[a + t - \tau_1 \le \tau_2 \middle| s_2\right] \ge w_2; \tag{11}$$

- The MNF at stress level s_3 is given as w_3 :

$$(n - n_1 - n_2)\Pr\left[b + t - \tau_2 \le c \middle| s_3\right] \ge w_3, \tag{12}$$

where n_i is the number of failures under each stress level s_i, $i = 1, 2, 3$.

With these constraints, the optimal decision variables (τ_1, τ_2, s_2) can be determined by minimizing each of the above-mentioned loss functions. Specifically, we choose the optimal designs in order to minimize (7), (8), or (9) with respect to (τ_1, τ_2 , s_2) subject to all three constraints (10), (11), and (12). We denote the optimal designs corresponding to the three optimality criteria by ${}^D\xi$, ${}^A\xi$, and ${}^Q\xi$, which can be expressed as

$${}^D\xi = \arg\{\min(L_D \mid \text{given (10), (11), and (12)})\}, \tag{13}$$

$${}^A\xi = \arg\{\min(L_A \mid \text{given (10), (11), and (12)})\}, \tag{14}$$

and

$${}^Q\xi = \arg\{\min(L_Q \mid \text{given (10), (11), and (12)})\}. \tag{15}$$

3.3. Optimal designs when the middle stress being fixed

In [10], Elsayed and Zhang discussed an example of conducting a two-step-stress ALT experiment for metal oxide semiconductor capacitors and estimating the hazard rate over 10 years at design temperature 50^oC. The total number of test units was n = 200 and the test was censored by c = 300 hours. In order to avoid any unanticipated change in failure mechanisms during the ALT experiment, the maximum testing temperature was determined to be 250^oC by the engineering experimenters. The initial values of the model parameters was taken as $\gamma_0 = 0.0001$, $\gamma_1 = 0.5$, and $\beta = -3800$. The Q-optimal low accelerated stress level was found to be 145^oC.

In this subsection, we use this example to compare the optimal two-step-stress ALT designs obtained in [10] with our proposed three-step-stress ALT designs. Previously, the optimal two-step-stress ALT designs are having $s_1 = 145^oC$ and $s_2 = 250^oC$. For our three-step-stress ALT plans, we keep the number of test units, censoring time, and the lowest and highest stress levels all the same. Namely, n = 200, c = 300 , $s_1 = 145^oC$ and $s_3 = 250^oC$. The initial values for the model parameters are kept the same as well. Conveniently, the experimenters often take s_2 as the average of s_1 and s_3. Therefore, we discuss two scenarios for the choice of s_2: (a) being 197.5^oC, which is the average of $s_1 = 145^oC$ and $s_3 = 250^oC$; and (b) being optimally chosen, together with stress changing times, in order to minimize a specified loss function. We note that when temperature is taken as a stress factor appeared in a PH model for ALT, absolute temperature is often used as its measurement unit for model fitting.

Table 1. Constraint cases for two-step-stress ALT

Notation	Constraint parameters	MNF
C_{12}	$w_1 = 40$, $w_2 = 30$	70
C_{34}	$w_1 = 30$, $w_2 = 40$	70
C_{55}	$w_1 = 40$, $w_2 = 20$	60
C_{66}	$w_1 = 20$, $w_2 = 40$	60

Table 2. Constraint cases for three-step-stress ALT

Notation	Constraint parameters	MNF
C_1	$w_1 = 40$, $w_2 = 20$, $w_3 = 10$	70
C_2	$w_1 = 40$, $w_2 = 15$, $w_3 = 15$	70
C_3	$w_1 = 30$, $w_2 = 30$, $w_3 = 10$	70
C_4	$w_1 = 30$, $w_2 = 20$, $w_3 = 20$	70
C_5	$w_1 = 40$, $w_2 = 10$, $w_3 = 10$	60
C_6	$w_1 = 20$, $w_2 = 20$, $w_3 = 20$	60

In Scenario (a), the optimal stress changing times, τ_1 and τ_2, can be chosen in order to minimize L_D, L_A, or L_Q under some specific constraints. We take several practical constraint cases with given w_1, w_2 and w_3, and these cases accompanied by their notations are recorded in Table 1 for two-step-stress ALT designs, and in Table 2 for three-step-stress ALT designs.

We denote the D-, A- and Q-optimal designs obtained for two-step-stress ALT under a given constraint case C_k by ${}^{D}\xi^{2}_{C_k}$, ${}^{A}\xi^{2}_{C_k}$, and ${}^{Q}\xi^{2}_{C_k}$, where $k = 12, 34, 55, 66$, and the D-, A- and Q-optimal designs obtained for three-step-stress ALT under a given constraint C_k by ${}^{D}\xi^{3(i)}_{C_k}$, ${}^{A}\xi^{3(i)}_{C_k}$, and ${}^{Q}\xi^{3(i)}_{C_k}$, where $k = 1, \ldots, 6$, and $i = 1, 2$ with i referring to the different scenario of s_i ($i = 1$ for Scenario (a), and $i = 2$ for Scenario (b)). Moreover, we define the asymptotic D-, A-, Q-efficiencies of $\xi^{3(i)}_{C_k}$ relative to $\xi^{2}_{C_k}$ as

$$
\begin{aligned}
{}^{D}eff\left(3(i),2\right) &= \frac{L_D\left({}^{D}\xi^{2}_{C_k}\right)}{L_D\left({}^{D}\xi^{3(i)}_{C_k}\right)},\\
{}^{A}eff\left(3(i),2\right) &= \frac{L_A\left({}^{A}\xi^{2}_{C_k}\right)}{L_A\left({}^{A}\xi^{3(i)}_{C_k}\right)}, \text{ and}\\
{}^{Q}eff\left(3(i),2\right) &= \frac{L_Q\left({}^{Q}\xi^{2}_{C_k}\right)}{L_Q\left({}^{Q}\xi^{3(i)}_{C_k}\right)}, \text{ for } i = 1,\ 2;
\end{aligned}
$$

and the asymptotic D-, A-, Q-efficiencies of $\xi^{3(i)}_{C_k}$ relative to $\xi^{3(j)}_{C_k}$ as

$$
\begin{aligned}
{}^{D}eff\left(3(i),3(j)\right) &= \frac{L_D\left({}^{D}\xi^{3(j)}_{C_k}\right)}{L_D\left({}^{D}\xi^{3(i)}_{C_k}\right)},\\
{}^{A}eff\left(3(i),3(j)\right) &= \frac{L_A\left({}^{A}\xi^{3(j)}_{C_k}\right)}{L_A\left({}^{A}\xi^{3(i)}_{C_k}\right)}, \text{ and}\\
{}^{Q}eff\left(3(i),3(j)\right) &= \frac{L_Q\left({}^{Q}\xi^{3(j)}_{C_k}\right)}{L_Q\left({}^{Q}\xi^{3(i)}_{C_k}\right)}, \text{ for } i, j = 1,\ 2.
\end{aligned}
$$

The optimal stress-changing time(s), for both two-step-stress plans and the three-step-stress ALT plans with s_2 being fixed at $197.5^{o}C$, are obtained by minimizing L_D, L_A, or L_Q, respectively for all the constraint cases considered. We note that the resulting optimal stress-changing times are all the same for these three different optimal criteria although their corresponding relative efficiencies are not the same. The optimal stress-changing times and relative efficiencies for all six constraint cases (as listed in Table 2) are presented in Tables 3 and 4. The overall average efficiency gain after unitizing an optimal three-step-stress plan is 5.44% when s_2 being fixed at $197.5^{o}C$.

4. Optimal designs when baseline hazard is a quadratic function

4.1. Preliminary

We have considered the PH based ALT with the case where the baseline hazard function is a simple linear function in Section 3. In many practical situations, this simple model can be under-fitted. Therefore, in this section, we focus on the PH model when the baseline hazard function is in a quadratic form. Namely, we adopt Model (m1) where the baseline hazard function is rather being

$$\lambda_0(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2. \tag{16}$$

Table 3. D-, A-, Q-optimal stress-changing times, when $s_2 = 197.5^{o}C$

3-step-stress ALT			2-step-stress ALT	
optimal designs	τ_1	τ_2	optimal designs	τ_1
${}^{D,A,Q}\xi^{3(1)}_{C_1}$	178.2	223.6	${}^{D,A,Q}\xi^{2}_{C_{12}}$	183
${}^{D,A,Q}\xi^{3(1)}_{C_2}$	178.0	209.2	${}^{D,A,Q}\xi^{2}_{C_{12}}$	183
${}^{D,A,Q}\xi^{3(1)}_{C_3}$	162.9	219.4	${}^{D,A,Q}\xi^{2}_{C_{34}}$	156
${}^{D,A,Q}\xi^{3(1)}_{C_4}$	162.1	194.8	${}^{D,A,Q}\xi^{2}_{C_{34}}$	156
${}^{D,A,Q}\xi^{3(1)}_{C_5}$	198.0	227.5	${}^{D,A,Q}\xi^{2}_{C_{55}}$	201
${}^{D,A,Q}\xi^{3(1)}_{C_6}$	162.1	194.8	${}^{D,A,Q}\xi^{2}_{C_{66}}$	156

Table 4. D-, A-, Q- relative efficiencies, when $s_2 = 197.5^{o}C$

	C_1	C_2	C_3	C_4	C_5	C_6
${}^{D}eff\left(3(1),2\right)$	1.06	1.04	1.10	1.06	1.04	1.06
${}^{A}eff\left(3(1),2\right)$	1.06	1.03	1.09	1.06	1.04	1.06
${}^{Q}eff\left(3(1),2\right)$	1.04	1.03	1.08	1.05	1.03	1.05

Consequently, the cdf, pdf, the cumulative hazard function, and the reliability function of failure time t, at a given stress level s, become

$$
\begin{aligned}
F(t;s) &= 1 - \exp\left[-\left(\gamma_0 t + \frac{\gamma_1}{2}t^2 + \frac{\gamma_2}{3}t^3\right)\exp(\beta s)\right],\\
f(t;s) &= \lambda_0(t)\exp(\beta s)\exp\left[-\left(\gamma_0 t + \frac{\gamma_1}{2}t^2 + \frac{\gamma_2}{3}t^3\right)\exp(\beta s)\right],\\
\Lambda(t;s) &= \left(\gamma_0 t + \frac{\gamma_1}{2}t^2 + \frac{\gamma_2}{3}t^3\right)\exp(\beta s), \text{ and}\\
R(t;s) &= \exp\left[-\left(\gamma_0 t + \frac{\gamma_1}{2}t^2 + \frac{\gamma_2}{3}t^3\right)\exp(\beta s)\right].
\end{aligned}
\tag{17}
$$

Then, the log-likelihood function of t, under a three-step-stress design ξ with stress levels being s_1, s_2, and s_3, become

$$
\begin{aligned}
\ln L(t;\xi) = {} & I_1 I_2 I_3\left[\ln\left(\gamma_0 + \gamma_1 t + \gamma_2 t^2\right) + \beta s_1 - \left(\gamma_0 t + \frac{\gamma_1}{2}t^2 + \frac{\gamma_2}{3}t^3\right)\exp(\beta s_1)\right]\\
& + (1 - I_1) I_2 I_3\left[\ln\left(\gamma_0 + \gamma_1 x + \gamma_2 x^2\right) + \beta s_2 - \left(\gamma_0 x + \frac{\gamma_1}{2}x^2 + \frac{\gamma_2}{3}x^3\right)\exp(\beta s_2)\right]\\
& + (1 - I_2) I_3\left[\ln\left(\gamma_0 + \gamma_1 y + \gamma_2 y^2\right) + \beta s_3 - \left(\gamma_0 y + \frac{\gamma_1}{2}y^2 + \frac{\gamma_2}{3}y^3\right)\exp(\beta s_3)\right]\\
& - (1 - I_3)\left[\left(\gamma_0 d + \frac{\gamma_1}{2}d^2 + \frac{\gamma_2}{3}d^3\right)\exp(\beta s_3)\right],
\end{aligned}
$$

where x, y, and d are defined as in (2). Its first partial derivatives with respect to the model parameters γ_0, γ_1, and β are the same as in (3) but with $\lambda_0(t)$ being defined as in (16) instead, and the one with respect to γ_2 is

$$\frac{\partial \ln L}{\partial \gamma_2} = I_1 I_2 I_3 t^2 \left[\frac{1}{\lambda_0(t)} - \frac{t}{3}\exp(\beta s_1)\right] + (1-I_1) I_2 I_3 x^2 \left[\frac{1}{\lambda_0(x)} - \frac{x}{3}\exp(\beta s_2)\right] + (1-I_2) I_3 y^2 \left[\frac{1}{\lambda_0(y)} - \frac{y}{3}\exp(\beta s_3)\right] - (1-I_3)\frac{d^3}{3}\exp(\beta s_3).$$

Table 5. D-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^{D}eff(3(2),2)$	${}^{D}eff(3(2),3(1))$
${}^{D}\xi^{3(2)}_{C_1}$	152.65	247.44	1.13	1.07
${}^{D}\xi^{3(2)}_{C_2}$	148.33	226.85	1.09	1.05
${}^{D}\xi^{3(2)}_{C_3}$	152.93	247.44	1.19	1.09
${}^{D}\xi^{3(2)}_{C_4}$	143.25	211.74	1.12	1.06
${}^{D}\xi^{3(2)}_{C_5}$	179.40	245.78	1.08	1.04
${}^{D}\xi^{3(2)}_{C_6}$	161.43	211.03	1.11	1.05

Table 6. A-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			A-relative efficiency	
Optimal design	τ_1	τ_2	${}^{A}eff(3(2),2)$	${}^{A}eff(3(2),3(1))$
${}^{A}\xi^{3(2)}_{C_1}$	155.20	247.47	1.13	1.07
${}^{A}\xi^{3(2)}_{C_2}$	147.99	226.84	1.09	1.06
${}^{A}\xi^{3(2)}_{C_3}$	155.13	247.47	1.19	1.09
${}^{A}\xi^{3(2)}_{C_4}$	142.41	211.73	1.11	1.05
${}^{A}\xi^{3(2)}_{C_5}$	155.18	247.47	1.09	1.05
${}^{A}\xi^{3(2)}_{C_6}$	142.39	211.73	1.11	1.05

Table 7. Q-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^{Q}eff(3(2),2)$	${}^{Q}eff(3(2),3(1))$
${}^{Q}\xi^{3(2)}_{C_1}$	156.19	247.47	1.09	1.05
${}^{Q}\xi^{3(2)}_{C_2}$	149.81	226.86	1.07	1.04
${}^{Q}\xi^{3(2)}_{C_3}$	156.71	247.47	1.14	1.06
${}^{Q}\xi^{3(2)}_{C_4}$	144.20	211.75	1.09	1.04
${}^{Q}\xi^{3(2)}_{C_5}$	156.45	247.47	1.06	1.03
${}^{Q}\xi^{3(2)}_{C_6}$	143.98	211.75	1.09	1.04

Its second partial derivatives $\frac{\partial^2 \ln L}{\partial \gamma_0^2}$, $\frac{\partial^2 \ln L}{\partial \gamma_1^2}$, $\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1}$, and $\frac{\partial^2 \ln L}{\partial \beta^2}$ can be expressed as the same as in Xu and Huang (2018) but with their λ_0 and Λ function being defined as in (16) and (17), and their corresponding elements of the Fisher's information matrix for a single failure time t (the expected value of negative second derivatives) are as the same as in (3), (4), (5), and (6). The remaining second derivatives are

$$\frac{\partial^2 \ln L}{\partial \gamma_2^2} \approx -\frac{I_1 I_2 I_3 t^4}{\lambda_0^2(t)} - \frac{(1-I_1) I_2 I_3 x^4}{\lambda_0^2(x)} - \frac{(1-I_2) I_3 y^4}{\lambda_0^2(y)},$$

$$\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2} \approx -\frac{I_1 I_2 I_3 t^2}{\lambda_0^2(t)} - \frac{(1-I_1) I_2 I_3 x^2}{\lambda_0^2(x)} - \frac{(1-I_2) I_3 y^2}{\lambda_0^2(y)},$$

$$\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2} \approx -\frac{I_1 I_2 I_3 t^3}{\lambda_0^2(t)} - \frac{(1-I_1) I_2 I_3 x^3}{\lambda_0^2(x)} - \frac{(1-I_2) I_3 y^3}{\lambda_0^2(y)},$$

and their corresponding elements of the Fisher's information matrix for a single t can be derived as below:

$$F_{22} = E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_2^2}\right\} = \int_0^{\tau_1} \frac{t^4}{\lambda_0^2(t)} f(t;s_1)\,dt + \int_a^{a+\tau_2-\tau_1} \frac{x^4}{\lambda_0^2(x)} f(x;s_2)\,dx + \int_b^{b+c-\tau_2} \frac{y^4}{\lambda_0^2(y)} f(y;s_3)\,dy$$
$$= \exp(\beta s_1)\int_0^{\tau_1} \frac{t^4}{\lambda_0(t)} R(t;s_1)\,dt + \exp(\beta s_2)\int_a^{a+\tau_2-\tau_1} \frac{x^4}{\lambda_0(x)} R(x;s_2)\,dx + \exp(\beta s_3)\int_b^{b+c-\tau_2} \frac{y^4}{\lambda_0(y)} R(y;s_3)\,dy,$$

$$F_{02} = E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2}\right\} = \int_0^{\tau_1} \frac{t^2}{\lambda_0^2(t)} f(t;s_1)\,dt + \int_a^{a+\tau_2-\tau_1} \frac{x^2}{\lambda_0^2(x)} f(x;s_2)\,dx + \int_b^{b+c-\tau_2} \frac{y^2}{\lambda_0^2(y)} f(y;s_3)\,dy$$
$$= \exp(\beta s_1)\int_0^{\tau_1} \frac{t^2}{\lambda_0(t)} R(t;s_1)\,dt + \exp(\beta s_2)\int_a^{a+\tau_2-\tau_1} \frac{x^2}{\lambda_0(x)} R(x;s_2)\,dx + \exp(\beta s_3)\int_b^{b+c-\tau_2} \frac{y^2}{\lambda_0(y)} R(y;s_3)\,dy,$$

and

$$F_{12} = E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2}\right\} = \int_0^{\tau_1} \frac{t^3}{\lambda_0^2(t)} f(t;s_1)\,dt + \int_a^{a+\tau_2-\tau_1} \frac{x^3}{\lambda_0^2(x)} f(x;s_2)\,dx + \int_b^{b+c-\tau_2} \frac{y^3}{\lambda_0^2(y)} f(y;s_3)\,dy$$
$$= \exp(\beta s_1)\int_0^{\tau_1} \frac{t^3}{\lambda_0(t)} R(t;s_1)\,dt + \exp(\beta s_2)\int_a^{a+\tau_2-\tau_1} \frac{x^3}{\lambda_0(x)} R(x;s_2)\,dx + \exp(\beta s_3)\int_b^{b+c-\tau_2} \frac{y^3}{\lambda_0(y)} R(y;s_3)\,dy,$$

with R function being defined as in (17).

As indicated in Section 3.1, the correlations between the stress coefficient β and baseline parameters γ_0, γ_1 and γ_2 are equal to zero. Consequently, the covariance matrix $\boldsymbol{\Sigma}$, of the MLEs of $\gamma_0, \gamma_1, \gamma_2, \beta$ can be expressed as

$$\boldsymbol{\Sigma} = \mathbf{F}^{-1} = \begin{bmatrix} Var\left(\mathring{\gamma}_0\right) & Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Var\left(\grave{\gamma}_1\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & Var\left(\mathring{\gamma}_2\right) & 0 \\ 0 & 0 & 0 & Var\left(\grave{\beta}\right) \end{bmatrix},$$

where $\mathbf{F}$ is the Fisher's information matrix of the full sample with size n and

$$\mathbf{F} = n\begin{bmatrix} E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0^2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2}\right\} & 0 \\ E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_1}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1^2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2}\right\} & 0 \\ E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_0 \gamma_2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_1 \gamma_2}\right\} & E\left\{-\frac{\partial^2 \ln L}{\partial \gamma_2^2}\right\} & 0 \\ 0 & 0 & 0 & E\left\{-\frac{\partial^2 \ln L}{\partial \beta^2}\right\} \end{bmatrix} = n\begin{bmatrix} F_{00} & F_{01} & F_{02} & 0 \\ F_{01} & F_{11} & F_{12} & 0 \\ F_{02} & F_{12} & F_{22} & 0 \\ 0 & 0 & 0 & F_\beta \end{bmatrix}.$$

4.2. Loss functions and design constraints

Now the loss functions under D-, A-, and Q-optimality become

$$\begin{aligned} L_D &= \det n \begin{bmatrix} Var\left(\mathring{\gamma}_0\right) & Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Var\left(\grave{\gamma}_1\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & Var\left(\mathring{\gamma}_2\right) & 0 \\ 0 & 0 & 0 & Var\left(\grave{\beta}\right) \end{bmatrix} \\ &= \det\left(\begin{bmatrix} F_{00} & F_{01} & F_{02} & 0 \\ F_{01} & F_{11} & F_{12} & 0 \\ F_{02} & F_{12} & F_{22} & 0 \\ 0 & 0 & 0 & F_\beta \end{bmatrix}\right)^{-1} \\ &= \frac{1}{\left(F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2\right)F_\beta}, \end{aligned} \tag{18}$$

$$\begin{aligned} L_A &= \mathbf{tr} n \begin{bmatrix} Var\left(\mathring{\gamma}_0\right) & Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \grave{\gamma}_1\right) & Var\left(\grave{\gamma}_1\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & 0 \\ Cov\left(\mathring{\gamma}_0, \mathring{\gamma}_2\right) & Cov\left(\grave{\gamma}_1, \mathring{\gamma}_2\right) & Var\left(\mathring{\gamma}_2\right) & 0 \\ 0 & 0 & 0 & Var\left(\grave{\beta}\right) \end{bmatrix} \\ &= \mathbf{tr}\left(\begin{bmatrix} F_0 & F_{01} & F_{02} & 0 \\ F_{01} & F_1 & F_{12} & 0 \\ F_{02} & F_{12} & F_2 & 0 \\ 0 & 0 & 0 & F_\beta \end{bmatrix}^{-1}\right) \\ &= \frac{F_{00}F_{11} + F_{00}F_{22} + F_{11}F_{22} - F_{01}^2 - F_{02}^2 - F_{12}^2}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} + \frac{1}{F_\beta}, \end{aligned} \tag{19}$$

and

$$\begin{aligned} L_Q &= \exp\left(-2\beta s_D\right)\int_0^T n Var\left[\grave{\lambda}\left(t; s_D\right)\right]dt \\ &= \exp\left(-2\beta s_D\right)\int_0^T n Var\left[\left(\mathring{\gamma}_0 + \grave{\gamma}_1 t + \mathring{\gamma}_2 t^2\right)\exp\beta s_D\right]dt \\ &= \exp\left(-2\beta s_D\right)\int_0^T \left[\frac{\partial\lambda}{\partial\gamma_0} \quad \frac{\partial\lambda}{\partial\gamma_1} \quad \frac{\partial\lambda}{\partial\gamma_2} \quad \frac{\partial\lambda}{\partial\beta}\right]\mathbf{F}^{-1}\left[\frac{\partial\lambda}{\partial\gamma_0} \quad \frac{\partial\lambda}{\partial\gamma_1} \quad \frac{\partial\lambda}{\partial\gamma_2} \quad \frac{\partial\lambda}{\partial\beta}\right]^T dt \\ &= \int_0^T \left[\frac{\left(F_{11}F_{22} - F_{12}^2\right) - 2t\left(F_{01}F_{22} - F_{02}F_{12}\right) + t^2\left(2F_{01}F_{12} - 2F_{11}F_{02} + F_{00}F_{22} - F_{02}^2\right) - 2t^3\left(F_{00}F_{12} - F_{01}F_{02}\right) + t^4\left(F_{00}F_{11} - F_{01}^2\right)}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} + \frac{s_D^2\left(\gamma_0 + \gamma_1 t + \gamma_2 t^2\right)^2}{F_\beta}\right]dt \\ &= \frac{\left(F_{11}F_{22} - F_{12}^2\right)T - \left(F_{01}F_{22} - F_{02}F_{12}\right)T^2 + \left(2F_{01}F_{12} - 2F_{11}F_{02} + F_{00}F_{22} - F_{02}^2\right)T^3}{F_{00}F_{11}F_{22} + 2F_{01}F_{02}F_{12} - F_{00}F_{12}^2 - F_{11}F_{02}^2 - F_{22}F_{01}^2} \\ &\quad \frac{-\frac{1}{2}\left(F_{00}F_{12} - F_{01}F_{02}\right)T^4 + \frac{1}{5}\left(F_{00}F_{11} - F_{01}^2\right)T^5}{} \\ &\quad + \frac{s_D^2\left(\frac{\gamma_1^2}{2}T + \gamma_0\gamma_1 T^2 + \frac{\gamma_1^2 + 2\gamma_0\gamma_2}{3}T^3 + \frac{\gamma_1\gamma_2}{2}T^4 + \frac{\gamma_2^2}{5}T^5\right)}{F_\beta}, \end{aligned} \tag{20}$$

respectively.

The optimal decision variables (τ_1 , τ_2 , s_2) are chosen by minimizing the loss function (18), (19), or (20) with some design constraints. We also keep the three constrains (10), (11), and (12) as the same as those in Section 3. The corresponding designs ${}^{D}\xi$, ${}^{A}\xi$ and ${}^{Q}\xi$ can also be described as (13), (14), and (15).

4.3. Optimal three-step-stress designs

In this subsection, we revisit the example presented in Section 3, and discuss the optimal ALT designs when the fitting model is a PH model with a quadratic baseline hazard function. Suppose from previous experience, the initial values for the model parameters are $\gamma_0 = 0.0001$, $\gamma_1 = 0.5$, $\gamma_2 = 0$, $\beta = -3800$. All other values of the design parameters in the example remain the same. Thus, the accelerated stress levels remain as $s_1 = 145^o C$, $s_3 = 250^o C$. Here we also consider the six constraint cases as used in Section 3.

When s_2 is conveniently chosen as $197.5^o C$, the resulting two-step-stress and three-step-stress optimal designs under D-, A, and Q-optimality criteria appeared to be the same again. Nonetheless, the relative efficiencies are not always better than their corresponding peers, the optimal two-step-stress designs. These designs and their relative efficiencies under D-, A-, and Q-optimalities are displayed in Tables 8 and 9, respectively.

Table 8. D-, A-, Q-optimal stress-changing times, when $s_2 = 197.5^o C$

3-step-stress ALT			2-step-stress ALT	
optimal designs	τ_1	τ_2	optimal designs	τ_1
${}^{D,A,Q}\xi_{C_1}^{3(1)}$	178.2	223.6	${}^{D,A,Q}\xi_{C_{12}}^{2}$	183
${}^{D,A,Q}\xi_{C_2}^{3(1)}$	178.0	209.2	${}^{D,A,Q}\xi_{C_{12}}^{2}$	183
${}^{D,A,Q}\xi_{C_3}^{3(1)}$	162.9	219.4	${}^{D,A,Q}\xi_{C_{34}}^{2}$	156
${}^{D,A,Q}\xi_{C_4}^{3(1)}$	162.1	194.8	${}^{D,A,Q}\xi_{C_{34}}^{2}$	156
${}^{D,A,Q}\xi_{C_5}^{3(1)}$	198.0	227.5	${}^{D,A,Q}\xi_{C_{55}}^{2}$	201
${}^{D,A,Q}\xi_{C_6}^{3(1)}$	162.1	194.8	${}^{D,A,Q}\xi_{C_{66}}^{2}$	156

Table 9. D-, A-, Q- relative efficiencies, when $s_2 = 197.5^o C$

	C_1	C_2	C_3	C_4	C_5	C_6
${}^{D}eff\left(3(1),2\right)$	0.93	0.89	1.14	1.0	0.94	1.0
${}^{A}eff\left(3(1),2\right)$	1.06	1.03	1.09	1.06	1.04	1.06
${}^{Q}eff\left(3(1),2\right)$	0.88	0.86	1.04	0.94	0.91	0.94

Table 10. D-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			D-relative efficiency	
Optimal design	τ_1	τ_2	${}^{D}eff\left(3(2),2\right)$	${}^{D}eff\left(3(2),3(1)\right)$
${}^{D}\xi^{3(2)}_{C_1}$	178.05	245.70	1.91	2.05
${}^{D}\xi^{3(2)}_{C_2}$	175.60	220.01	1.42	1.60
${}^{D}\xi^{3(2)}_{C_3}$	145.65	247.10	2.80	2.46
${}^{D}\xi^{3(2)}_{C_4}$	154.65	210.94	1.79	1.79
${}^{D}\xi^{3(2)}_{C_5}$	179.37	245.61	1.47	1.90
${}^{D}\xi^{3(2)}_{C_6}$	154.65	210.94	1.79	1.79

Table 11. A-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			A-relative efficiency	
Optimal design	τ_1	τ_2	${}^{A}eff\left(3(2),2\right)$	${}^{A}eff\left(3(2),3(1)\right)$
${}^{A}\xi^{3(2)}_{C_1}$	155.20	247.47	1.13	1.07
${}^{A}\xi^{3(2)}_{C_2}$	147.99	226.84	1.09	1.06
${}^{A}\xi^{3(2)}_{C_3}$	155.13	247.47	1.19	1.09
${}^{A}\xi^{3(2)}_{C_4}$	142.41	211.73	1.11	1.05
${}^{A}\xi^{3(2)}_{C_5}$	155.18	247.47	1.09	1.05
${}^{A}\xi^{3(2)}_{C_6}$	142.39	211.73	1.11	1.05

Very similar to the results found in Section 3 when the baseline hazard function being simple linear, the resulting designs with an optimal middle stress level provide much more efficiency gains than those with a fixed middle stress level when the baseline hazard function being quadratic. The optimal stress-changing times and the optimal middle stress levels, under D-, A, and Q-optimality criteria, for Model (1) with (16) are displayed in Tables 10, 11, and 12, respectively. These tables also include the efficiencies of the resulting optimal three-step-stress designs relative to both their corresponding optimal two-step-stress designs and optimal three-step-stress designs listed in Table 8, under D-, A-, and Q-optimalities.

Table 12. Q-optimal designs and relative efficiencies

3-step-stress ALT, $s_2 = 155$			Q-relative efficiency	
Optimal design	τ_1	τ_2	${}^{Q}eff\left(3(2),2\right)$	${}^{Q}eff\left(3(2),3(1)\right)$
${}^{Q}\xi^{3(2)}_{C_1}$	178.15	245.96	1.71	1.94
${}^{Q}\xi^{3(2)}_{C_2}$	177.96	224.53	1.39	1.62
${}^{Q}\xi^{3(2)}_{C_3}$	148.51	247.29	2.36	2.27
${}^{Q}\xi^{3(2)}_{C_4}$	162.04	210.97	1.62	1.72
${}^{Q}\xi^{3(2)}_{C_5}$	198.03	241.97	1.37	1.51
${}^{Q}\xi^{3(2)}_{C_6}$	162.04	210.97	1.62	1.72

We note that these three-step-stress ALT designs with optimal middle stress level have largely reduced the loss function for all cases. From Tables 10-12, the efficiency gains among all the cases considered are of a minimum of 9% and a maximum of 180% with respect to the optimal two-step-stress designs. The overall average efficiency gain is as high as 55.4% over the optimal two-step-stress designs and 59.7% over the optimal three-step-stress designs with a fixed middle stress. Such efficiency gains are much higher than the efficiency gains when adopting a PH model with a simple linear baseline function. The resulting optimal stress-changing times vary under three criteria, but all the optimal middle stress levels are equal being the lower bound of s_2. For all three criteria, the most efficient design occurs when the constraint case is C_3. Therefore, if the experimenter is uncertain which constraints should be applied, we would recommend to apply optimal designs ${}^{D}\xi^{3(2)}_{C_3}$ or ${}^{A}\xi^{3(2)}_{C_3}$ for better model parameter estimation and ${}^{Q}\xi^{3(2)}_{C_3}$ for more accurate hazard rate prediction.

5. Simulations

In order to demonstrate the performance of resulting designs obtained in Sections 3 and 4 with a given sample size, we carry out a simulation study. We first provide a procedure to simulate data from a given step-stress ALT experimental design, then we examine and compare the performances of the optimal designs constructed by the previous sections.

5.1. A simulation procedure

The following four steps describe our simulation procedure:

Step 1: Generating the number of failures and the failure times under the low stress level, s_1:

Given the initial values of true model parameters with the stress levels and stress-changing times of an optimal design, we can use a binomial distribution to simulate the data. We divide $[0,\tau_1]$, τ_1, into m_1 subintervals: $[t_{10},t_{11}]$, $(t_{11},t_{12}]$, ..., $(t_{1,m_1-1},t_{1m_1}]$, where

$0 = t_{10} < t_{11} < \cdots < t_{1m_1} = \tau_1$. Let X_i be the failure number over the i th subinterval $(t_{1,i-1}, t_{1i}]$, we assume that the distribution of X_i is a binomial distribution; namely,

$$X_i \sim B\left(n - \sum_{j=1}^{i-1} X_j,\ p_{1i}\right), i = 1, \ldots, m_1,$$

where p_{1i} is the failure rate within the ith interval $(t_{1,i-1}, t_{1i}]$ under the low stress level, s_1. Then, p_{1i} can be derived as follows:

$$p_{1i} = P\{t_{1,i-1} < t \le t_{1i} \mid s_1\} = R(t_{1,i-1}; s_1) - R(t_{1i}; s_1)$$
$$= \exp\left[-\left(\gamma_0 t_{1,i-1} + \frac{\gamma_1}{2} t_{1,i-1}^2 + \frac{\gamma_2}{3} t_{1,i-1}^3\right)\exp(\beta s_1)\right] - \exp\left[-\left(\gamma_0 t_{1i} + \frac{\gamma_1}{2} t_{1i}^2 + \frac{\gamma_2}{3} t_{1i}^3\right)\exp(\beta s_1)\right].$$

Note that when the baseline function is a simple linear function, we let $\gamma_2 = 0$. Then, we generate X_i failure times randomly from a uniform distribution within each subinterval $(t_{1,i-1}, t_{1i}]$. Therefore, the total number of failure times generated under s_1 is $n_1 = \sum_{i=1}^{m_1} X_i$.

Step 2: Generating the number of failures and the failure times under the middle stress level, s_2:

Similarly to Step 1, we divide the second time period $(\tau_1, \tau_2]$ into m_2 subintervals: $(t_{20}, t_{21}]$, $(t_{21}, t_{22}]$, ..., $(t_{2,m_2-1}, t_{2m_2}]$, where $\tau_1 = t_{20} < t_{21} < \cdots < t_{2m_2} = \tau_2$. Let Y_i be the failure number over the ith interval $(t_{2,i-1}, t_{2i}]$, the distribution of Y_i can be assumed as

$$Y_i \sim B\left((n - n_1) - \sum_{j=1}^{i-1} Y_j,\ p_{2i}\right), i = 1, \ldots, m_2,$$

where p_{2i} is the failure rate within the ith interval $(t_{2,i-1}, t_{2i}]$ under the middle stress level, s_2. Then, p_{2i} can be computed by

$$p_{2i} = P\{a + t_{2,i-1} < t \le a + t_{2i} \mid s_2\}$$
$$= \exp\left(-\left[\gamma_0\left(a + t_{2,i-1}\right) + \frac{\gamma_1}{2}\left(a + t_{2,i-1}\right)^2 + \frac{\gamma_2}{3}\left(a + t_{2,i-1}\right)^3\right]\exp(\beta s_2)\right)$$
$$- \exp\left(-\left[\gamma_0\left(a + t_{2i}\right) + \frac{\gamma_1}{2}\left(a + t_{2i}\right)^2 + \frac{\gamma_2}{3}\left(a + t_{2i}\right)^3\right]\exp(\beta s_2)\right),$$

where $a = F_2^{-1}\left[F_1(\tau_1)\right]$. Then, we generate Y_i failure times randomly from a uniform distribution within $(t_{2,i-1}, t_{2i}]$ for each i. Therefore, the total number of failure times generated under s_2 is $n_2 = \sum_{i=1}^{m_2} Y_i$.

Step 3: Generating the number of failures and the failure times under the high stress level, s_3:

Now, we divide the third time period $(\tau_2, c]$ into m_3 subintervals: $(t_{30}, t_{31}]$, $(t_{31}, t_{32}]$, ..., $(t_{3,m_3-1}, t_{3m_3}]$, where $\tau_2 = t_{30} < t_{31} < \cdots < t_{3m_3} = c$. Let Z_i be the failure number over the ith interval $(t_{3,i-1}, t_{3i}]$. The distribution of Z_i can be assumed as

$$Z_i \sim B\left((n - n_1 - n_2) - \sum_{k=1}^{i-1} Z_k,\ p_{3i}\right), i = 1, \ldots, m_3,$$

where p_{3i} is the failure rate within the ith interval $(t_{3,i-1}, t_{3i}]$ under the high stress level, s_3. Then, p_{3i} can be derived as

$$p_{3i} = P\{b + t_{3,i-1} < t \le b + t_{3i} \mid s_3\}$$
$$= \exp\left(-\left[\gamma_0\left(b + t_{3,i-1}\right) + \frac{\gamma_1}{2}\left(b + t_{3,i-1}\right)^2 + \frac{\gamma_2}{3}\left(b + t_{3,i-1}\right)^3\right]\exp(\beta s_3)\right)$$
$$- \exp\left(-\left[\gamma_0\left(b + t_{3i}\right) + \frac{\gamma_1}{2}\left(b + t_{3i}\right)^2 + \frac{\gamma_2}{3}\left(b + t_{3i}\right)^3\right]\exp(\beta s_3)\right),$$

where $b = F_3^{-1}\left[F_2\left(a + \tau_2 - \tau_1\right)\right]$. Then, we generate Z_i failure times randomly from a uniform distribution within $(t_{3,i-1}, t_{3i}]$ for each i, and finally the total number of failure times generated under s_3 is $n_3 = \sum_{i=1}^{m_3} Z_i$.

Step 4: Estimating the parameters:

For simplicity, we keep the length of all subinterval equal to q hours in this paper. The log-likelihood function can be expressed as

$$\begin{aligned} l(\boldsymbol{\beta}; \xi) &= n_1 \beta s_1 + \sum_{i=1}^{m_1}\left[X_i\left(\ln\left(\gamma_0 + \gamma_1 (qi)\right) - \left(\gamma_0 (qi) + \frac{\gamma_1}{2}(qi)^2\right)\exp(\beta s_1)\right)\right] \\ &+ n_2 \beta s_2 + \sum_{i=1}^{m_2}\left[Y_i\left(\ln\left(\gamma_0 + \gamma_1 (a + qi)\right) - \left(\gamma_0 (a + qi) + \frac{\gamma_1}{2}(a + qi)^2\right)\exp(\beta s_2)\right)\right] \\ &+ n_3 \beta s_3 + \sum_{i=1}^{m_3}\left[Z_i\left(\ln\left(\gamma_0 + \gamma_1 (b + qi)\right) - \left(\gamma_0 (b + qi) + \frac{\gamma_1}{2}(b + qi)^2\right)\exp(\beta s_3)\right)\right] \\ &- \left(n - n_1 - n_2 - n_3\right)\left[\left(\gamma_0 (b + c - \tau_2) + \frac{\gamma_1}{2}(b + c - \tau_2)^2\right)\exp(\beta s_3)\right]. \end{aligned} \quad (21)$$

For each simulation run, we can compute the maximum likelihood estimates of the model parameters by maximizing (21).

Let r be the number of simulation runs. We may use these r estimates to calculate the simulated squared bias (SBIAS^2), simulated variance (SVAR), and simulated mean squared error (SMSE) of each parameter estimator. We define SBIAS^2, SVAR, and SMSE for an estimator $\hat{\theta}$ as:

$$\text{SBIAS}^2\left(\hat{\theta}\right) = \left(\frac{1}{r}\sum_{j=1}^{r}\hat{\theta}_j - \theta_0\right)^2,$$
$$\text{SVAR}\left(\hat{\theta}\right) = \frac{1}{r-1}\sum_{j=1}^{r}\left(\hat{\theta}_j - \frac{1}{r}\sum_{j=1}^{r}\hat{\theta}_j\right)^2,$$
$$\text{and SMSE}\left(\hat{\theta}\right) = \frac{1}{r}\sum_{j=1}^{r}\left(\hat{\theta}_j - \theta_0\right)^2, \quad (22)$$

where θ_0 is the given true parameter value, $\hat{\theta}_j$ is the estimate of θ_0 from the jth simulation run. By these definitions, we can compute SBIAS^2, SVAR, and SMSE respectively for $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\beta}$ using the resulting designs from the example obtained in Sections 3 and 4. We can also compute SBIAS^2, SVAR, and SMSE of the MLE of hazard rate over T under normal design conditions for our resulting designs. The true hazard function over T under normal design stress level s_D can be expressed as

$$g(\boldsymbol{\alpha}) = \int_0^T \lambda(t; s_D)\, dt = \left(\gamma_0 T + \frac{\gamma_1}{2}T^2 + \frac{\gamma_2}{3}T^3\right)\exp(\beta s_D) \quad (23)$$

with $\boldsymbol{\alpha} = [\gamma_0, \gamma_1, \gamma_2, \beta]^T$. By the invariance property of MLEs, the MLE of (g1true) is

$$\hat{g} = \left(\hat{\gamma}_0 T + \frac{\hat{\gamma}_1}{2}T^2 + \frac{\hat{\gamma}_2}{3}T^3\right)\exp(\hat{\beta} s_D). \quad (24)$$

Then, SBIAS^2, SVAR, and SMSE of $\hat{g}$ can also be computed by (22).

5.2. Performance of the optimal designs obtained in Sections 3 and 4

Taking n = 200, c = 300 hours, and $r = 1000$, the length of all subintervals to be 5 hours, we use the simulation procedure introduced in Section 5.1 to evaluate the performance of our resulting designs. We adopt the constraint case C_1 for a demonstration.

For the PH model with a simple linear baseline function, from Section 3.3, ${}^{D,A,Q}\xi_{C_1}^{3(1)}$ having $[\tau_1 = 178.2,\ \tau_2 = 223.6,\ s_2 = 197.5]$, is the D-, A-, and Q-optimal design with fixed middle stress at $s_2 = 197.5^o C$. After we update s_2 to the optimal middle stress level, we get new D-, A-, and Q-optimal designs for C_1. Since the resulting D-, A- and Q-optimal designs are similar, we only present the simulation result for our Q-optimal design. The Q-optimal design we obtained in Section 3.4 for C_1 is ${}^{Q}\xi_{C_1}^{3(2)}$ with $[\tau_1 = 156.19,\ \tau_2 = 247.47,\ s_2 = 155]$. We denote the efficiencies of a design ξ_A relative to another design ξ_B in terms of SBIAS^2, SVAR, and SMSE as $eff_B(\xi_A, \xi_B)$, $eff_V(\xi_A, \xi_B)$, and $eff_M(\xi_A, \xi_B)$, respectively. Based on these 1000 simulation runs, we compute all the SBIAS^2, SVAR, and SMSE of the MLE for γ_0, γ_1, β, and g when each of the two different designs, ${}^{Q}\xi_{C_1}^{3(2)}$ and ${}^{Q}\xi_{C_1}^{3(1)}$, is adopted. Table 13 shows the efficiencies of ${}^{Q}\xi_{C_1}^{3(2)}$ relative to ${}^{Q}\xi_{C_1}^{3(1)}$ in terms of SBIAS^2, SVAR, and SMSE of $\hat{\beta}$, $\hat{\gamma}_0$, $\hat{\gamma}_1$, and (24) respectively. We note that since these are Q-optimal designs and therefore the efficiency gains appear more for estimating g. This is consistent with the previous finding of the asymptotic efficiency gains as discussed in Section 3. The simulation results indicate that the most efficiency gains appear in variance reduction with an extreme for γ_0. The reason behind is that the optimal designs constructed was aiming to minimize the asymptotic variances.

In Section 3, the Q-optimal designs are obtained by minimizing the asymptotic variance of the $\hat{g}$. As expected, when ${}^{Q}\xi_{C_1}^{3(2)}$ was adopted, we have gained the most efficiency (as high as 76.57%) in terms of the SVAR($\hat{g}$) compare to ${}^{Q}\xi_{C_1}^{3(1)}$. The SVAR of MLE of the model parameters are all very much reduced, and their efficiency gains of ${}^{Q}\xi_{C_1}^{3(2)}$ relative to ${}^{Q}\xi_{C_1}^{3(1)}$ in terms of SVAR($\hat{\gamma}_0$), SVAR($\hat{\gamma}_1$), and SVAR($\hat{\beta}$) are all higher than 100%. Moreover, SBIAS^2 and SMSE of $\hat{g}$ are also being reduced. All the results confirm that ${}^{Q}\xi_{C_1}^{3(2)}$ outperforms the design ${}^{Q}\xi_{C_1}^{3(1)}$.

Table 13. The efficiencies of ${}^{Q}\xi_{C_1}^{3(2)}$ relative to ${}^{Q}\xi_{C_1}^{3(1)}$ in terms of SBIAS^2, SVAR, and SMSE

	$\hat{\beta}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{g}$
$eff_B\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{Q}\xi_{C_1}^{3(1)}\right)$	0.9824	0.9803	1.0	1.0175
$eff_V\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{Q}\xi_{C_1}^{3(1)}\right)$	2.2207	24508114.61	2.0038	1.7657
$eff_M\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{Q}\xi_{C_1}^{3(1)}\right)$	0.9838	0.9851	1.0	1.0345

For the PH model with a quadratic baseline function, the optimal design ${}^{D,A,Q}\xi_{C_1}^{3(1)}$ $[\tau_1 = 178.15,\ \tau_2 = 223.57,\ s_2 = 197.5]$ (please see Table 8) is the D-, A, and Q-optimal designs with a fixed middle stress at $s_2 = 197.5^o C$. When we simultaneously choose an optimal middle stress together with optimal stress-changing times, the D-, A-, and Q-optimal designs become ${}^{D}\xi_{C_1}^{3(2)}$ having $[\tau_1 = 178.15, \tau_2 = 245.96]$, ${}^{A}\xi_{C_1}^{3(2)}$ having $[\tau_1 = 155.20,\ \tau_2 = 247.47]$, and ${}^{Q}\xi_{C_1}^{3(2)}$ having $[\tau_1 = 178.15, \tau_2 = 245.96]$, and they all are with $s_2 = 155$. We note that the D-optimal design and Q-optimal design are quite similar. Thus, we only present the simulation results for our resulting A- and Q-optimal designs in this example. Based on the 1000 simulation runs, we compute all SBIAS^2, SVAR, and SMSE of the MLE for γ_0, γ_1, γ_2, β and g when each of the three different designs, ${}^{D,A,Q}\xi_{C_1}^{3(1)}$, ${}^{A}\xi_{C_1}^{3(2)}$, and ${}^{Q}\xi_{C_1}^{3(2)}$, is adopted. Tables 14 and 15 display the efficiencies of ${}^{A}\xi_{C_1}^{3(2)}$ and ${}^{Q}\xi_{C_1}^{3(2)}$ relative to ${}^{D,A,Q}\xi_{C_1}^{3(1)}$ in terms of SBIAS^2, SVAR, and SMSE of $\hat{\beta}$, $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and (24) respectively. The simulation results have shown that the designs, ${}^{A}\xi_{C_1}^{3(2)}$ and ${}^{Q}\xi_{C_1}^{3(2)}$, by optimally selecting the middle stress level and stress-changing time simultaneously can reduce SVAR($\hat{g}$) by 282% and 183% compared to ${}^{D,A,Q}\xi_{C_1}^{3(1)}$. We also note that not all of SVARs of MLE of model parameters have been reduced much by using ${}^{A}\xi_{C_1}^{3(2)}$ or ${}^{Q}\xi_{C_1}^{3(2)}$. Only SVAR($\hat{\gamma}_0$) and SVAR($\hat{\gamma}_2$) have significantly lessened among all SVARs.

Further, we uncover that both SBIAS^2 and SMSE of all $\overset{\circ}{\gamma}_0$, $\overset{\prime}{\gamma}_1$, $\overset{\circ}{\gamma}_2$, $\overset{\prime}{\beta}$, and $\overset{\prime}{g}$ are reduced by adopting either ${}^{A}\xi_{C_1}^{3(2)}$ or ${}^{Q}\xi_{C_1}^{3(2)}$. The efficiencies of ${}^{A}\xi_{C_1}^{3(2)}$ or ${}^{Q}\xi_{C_1}^{3(2)}$ in terms of SMSE($\overset{\prime}{g}$) are more than 50 and 39 times higher than ${}^{D,A,Q}\xi_{C_1}^{3(1)}$. This indicates that unitizing ${}^{A}\xi_{C_1}^{3(2)}$ or ${}^{Q}\xi_{C_1}^{3(2)}$ can provide a great efficiency gain when experimenters are interested in estimating a hazard rate.

Conclusion

Optimal three-step-stress ALT designs for PH models, with either a linear or a quadratic baseline function, have been constructed in this paper. For a three-step-stress ALT, the practitioner often naturally set the average of high and low stress as the middle stress level. Nevertheless, from the results of both simulated and asymptotic efficiency comparison, we have revealed that the optimal three-step-stress ALT designs with both optimal stress-changing times and with an optimal middle stress level outperform the most among all the designs and for all the scenarios considered. Therefore, the three-step-stress plans with both optimal stress-changing times and optimal middle stress level are recommended especially when the hazard rate prediction is interested.

In Section 3, we have presented the resulting optimal designs for a practical ALT example when fitting a PH model with a simple linear baseline hazard function. Taking six different MNF constraint cases, we have found the optimal allocations of the stress-changing times and the optimal middle stress level that can minimize the loss function L_D, L_A, or L_Q. Thus, we have solved the minimization problem for a nonlinear objective function with multiple nonlinear constraints (MNF at different stress levels), and obtained the constrained optimal designs under each of D-, A-, and Q-optimality. The resulting optimal designs under three different criteria are quite similar. In addition, we have also found that the middle stress level should be kept as close to the lower bound of the middle stress level as possible as long as the constraint condition is satisfied.

Table 14. The efficiencies of ${}^{A}\xi_{C_1}^{3(2)}$ relative to ${}^{D,A,Q}\xi_{C_1}^{3(1)}$ in terms of SBIAS^2, SVAR, and SMSE

	$\overset{\prime}{\beta}$	$\overset{\circ}{\gamma}_0$	$\overset{\prime}{\gamma}_1$	$\overset{\circ}{\gamma}_2$	$\overset{\prime}{g}$
$eff_B\left({}^{A}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	1.0098	21.2561	1.0004	504.6733	313.2065
$eff_V\left({}^{A}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	0.9713 6	11.4924	0.0277	420.4643	2.8329
$eff_M\left({}^{A}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	1.0097	14.0173	1.0004	42.7565	40.2554

In Section 4, we have derived the optimal designs when fitting a PH model with a quadratic baseline hazard function. The optimal stress-changing times and the optimal middle stress level are also chosen in order to minimize the loss functions under given nonlinear constraints. Similarly to the case with a linear baseline hazard, under these constraints, the optimal middle stress level can be located as close to the low stress level as possible as long as such constraints are satisfied. For each of D-, A-, and Q-optimality, six different constrained optimal designs have been obtained. We also reveal that designing a three-step-stress ALT with a fixed middle stress seems only helping reduce the value of loss function under A-optimality. Thus, we suggest not conveniently taking the average of other two stress levels as the middle stress level, especially when a quadratic baseline hazard function is considered. We conclude that optimal three-step-stress designs have gained efficiency on an average of 68% for a hazard rate prediction compared to the corresponding optimal two-step-stress designs, which means that there is 7.56 times higher efficiency gain attained than the case when the baseline function is simple linear.

Table 15. The efficiencies of ${}^{Q}\xi_{C_1}^{3(2)}$ relative to ${}^{D,A,Q}\xi_{C_1}^{3(1)}$ in terms of SBIAS^2, SVAR, and SMSE

	$\overset{\prime}{\beta}$	$\overset{\circ}{\gamma}_0$	$\overset{\prime}{\gamma}_1$	$\overset{\circ}{\gamma}_2$	$\overset{\prime}{g}$
$eff_B\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	1.0271	31.4599	1.0004	199.0405	280.6279
$eff_V\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	0.7898	6.8388	0.0149	309.2855	3.8197
$eff_M\left({}^{Q}\xi_{C_1}^{3(2)}, {}^{D,A,Q}\xi_{C_1}^{3(1)}\right)$	1.0258	9.8674	1.0004	29.3142	51.1622

In Section 5, we have evaluated the performance of our resulting designs from both Sections 3 and 4 by simulations. The design for a three-step-stress ALT with an optimal middle stress level and two optimal stress-changing times has greatly increased the simulated efficiency of the hazard rate estimator. It is confirmed that the optimal designs with optimal middle stress significantly outperform those ones with the middle stress level fixed at the average of other two stress levels.

We note that although the design construction in this paper has been demonstrated by optimal designing three-step-stress ALT experiments when a single stress factor is involved, the method developed can be easily extended to conducting the optimal designs for more complicated multiple step-stress ALT, such as step-stress ALT with more than three steps, and/or ALT with multiple stress factors (but one of them engaged in conducting step-stress plans). We also notice that the inaccuracy of the assumed baseline hazard function can cause unavoidable prediction bias. Although the proposed designs perform very well when the model assumed is correct, they seem not helping much in reducing such bias when the assumed model is incorrect. If possible imprecision of the assumed PH model is suspected, then robust design approach should be considered. Some discussion on robust design for a PH model can be seen in [18].

Conflict of Interest

We declare that there is no any conflict of interest.

Acknowledgment

The research is supported by the Natural Sciences and Research Council of Canada.

References

[1] X. Xu, W. Huang, “Constrained Optimal Designs for Step-stress Accelerated Life Testing Experiments” *Proceedings of the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis*, 2018, 633-638.

[2] X. Xu, S. Hunt, "Robust Designs of Step-Stress Accelerated Life Testing Experiments for Reliability Prediction" *Matematika*, 29(1), 203-212, 2013.
[3] R. Miller, W. Nelson, "Optimum Simple Step-Stress Plans for Accelerated Life Testing" *IEEE Transactions in Reliability,* 32(1), 59-65, 1983.
[4] D. S. Bai, M. S. Kim, S. H. Lee, "Optimum simple step-stress accelerated life tests with censoring" *IEEE Transactions in Reliability,* 38(5), 528-532, 1989.
[5] D. S. Bai, M. S. Kim, "Optimum simple step-stress accelerated life tests for Weibull distribution and Type-I censoring" *Naval Research Logistics*, 40(2), 193-210, 1993.
[6] N. Fard, C. Li, "Optimal simple step stress accelerated life test design for reliability prediction" *Journal of Statistical Planning and Inference*, 139(5), 1799-1808, 2009.
[7] S. Hunt, X. Xu, "Optimal Design for Accelerated Life Testing with Simple Step-Stress Plans" *International Journal of Performability Engineering*, 8(5), 575-579, 2012.
[8] H. Ma, W. Q. Meeker, "Optimum step-stress accelerated life test plans for log-location-scale distributions" *Naval Research Logistics*, 55(6), 551-562, 2008.
[9] L. Jiao, "*Optimal allocations of stress levels and test units in accelerated life tests,*" Ph.D Thesis. Rutgers University-New Brunswick, New Brunswick, New Jersey, 2001.
[10] E. A. Elsayed, H. Zhang, "Design of Optimum Simple Step-Stress Accelerated Life Testing Plans" *Proceedings of the 2nd International IE Conference*, 1-14, 2004.
[11] C. H. Hu, R. D. Plante, J. Tang, "Step-stress accelerated life tests: a proportional hazards--based non-parametric model" *IIE Transactions*, 44(9) 754-764, 2012.
[12] N. Becker, B. McDonald, C. Khoo, "Optimal designs for fitting a proportional hazards regression model to data subject to censoring" Austral. J. Statist. 31, 449-468, 1989.
[13] H. Dette, M. Sahm, "Minimax optimal designs in nonlinear regression models" Statist. Sinica, 8, 1249-1264, 1998.
[14] J. M. McGree, J. A. Eccleston, "Investigating design for survival models" Metrika, 72, 295-311, 2010.
[15] J. L´opez-Fidalgo, M. J. Rivas-L´opez, R. Del Campo, "Optimal designs for Cox regression" Statistica Neerlandica 63, 135-148, 2009.
[16] M. Konstantinou, S. Biedermann, A. Kimber, "Optimal designs for two-parameter nonlinear models with application to survival models" *Statistica Sinica*, 24, 415-428, 2014.
[17] W. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*. Wiley. New York, 2004.
[18] X. Xu, W. Huang, "Optimal Robust Designs for Step-stress Accelerated Life Testing Experiments for Proportional Hazards Models" *Mathematical and Computational Approaches in Advancing Modern Science and Engineering*, Eds: Bélair, J., Frigaard, I.A., Kunze, H., Makarov, R., Melnik, R., Spiteri, R.J., New York: Springer, 585-594, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 119-126 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Assessment of Coordinated Multipoint Transmission Modes for Indoor and Outdoor Users at 28 GHz in Urban Macrocellular Environment

Muhammad Usman Sheikh[*,1,2], Ritayan Biswas[1],Jukka Lempiainen[1], Riku Jantti[2]

[1]*Tampere University, Faculty of Information Technology and Communication Sciences, Finland.*

[2]*Aalto University, Department of Communications and Networking, Finland.*

ARTICLE INFO

Article history:
Received: 04 December, 2018
Accepted: 03 March, 2019
Online: 16 March, 2019

Keywords:
CoMP
mmWave propagation
Macro cellular
3D ray tracing
Performance analysis

ABSTRACT

The aim of this article is to analyze and evaluate the performance of Coordinated Multipoint (CoMP) transmission approach at a frequency of 28 GHz using three dimensional ray tracing simulations in an urban macrocellular environment. The new performance metric introduced in this article is the relative power usage. Other performance metrics examined in this article are received power, the Signal to Interference plus Noise Ratio (SINR), user throughput, relative throughput gain, and the percentage of overlapping area with multiple cells. Indoor and outdoor users are separately analyzed for few key performance indicators. Different cases of coordinated multipoint transmission i.e. intra-node and inter-node coordination is analyzed. The post analysis of the acquired simulation data shows that the use of CoMP functionality is more beneficial for the cell edge users compared with the other users in terms of improving the user's experience. The throughput gain as well as the transmission overhead of the CoMP approach increases with the increase in number CoMP ports. Inter-node CoMP is much more power efficient and beneficial in comparison with the intra-node CoMP case.

1. Introduction

This article is an extension of the research work originally presented at Vehicular Technology Conference (VTC Spring 2018) [1]. In reference [1], the performance of Coordinated Multipoint Transmission (CoMP) was studied, and the performance metrics considered for the analysis were Signal to Interference plus Noise Ratio (SINR), throughput, and relative throughput gain. Whereas in this article the performance of CoMP transmission for the indoor and for the outdoor users is separately analyzed. Additional performance metrics included in this article are received signal strength, the percentage of carriers within the overlapping zone of 3 dB, 5 dB, and 7 dB, and relative power usage.

Cellular systems have evolved enormously since its inception. Current cellular technologies such as Long Term Evolution (LTE) and LTE-Advanced (LTE-A) are capable of providing data rates and capacity which are significantly higher than preceding technologies. However, the Fifth Generation (5G) of mobile communication has use cases, which have the requirement of thousand times more capacity in comparison with today's standards. This requirement is driven by bandwidth hungry applications, and those applications require high data rates. These applications may include uncompressed high definition video streaming, analyzing huge data sets for real time monitoring, and artificial intelligence applications to name a few. Additionally, the Cisco network index (VNI) report estimates that by the year 2020 there will be approximately 1.6 billion connected devices [2]. This indicates a huge increase in the number of connected devices and the amount of data utilized by these devices and applications. Therefore, in order to fulfill these requirements, different capacity enhancement solutions are needed. Several technologies such as network densification [3], heterogeneous networks, multi Carrier Aggregation (CA), Inter Cell Interference Coordination (ICIC), Enhanced ICIC (eICIC) [4], Coordinating Multipoint (CoMP) transmission and Millimeter wave (mmWave) communications [5-8] have been proposed for enhancing the capacity of cellular systems.

Traditionally, the cellular communication utilizes the microwave frequency band (sub 4 GHz frequency range) of the electromagnetic spectrum. Different operating frequency bands have been allocated for different mobile technologies. The mmWave frequency band, which lies in the range of 30 GHz – 300

[*] Muhammad Usman Sheikh, Email: muhammad.sheikh@tut.fi

GHz has the abundance of unused free spectrum. The large unlicensed bandwidth present at the mmWave frequencies can support huge data rates which is required in case of 5G technology. The mmWaves have certain attributes such as high path loss, high Building Penetration Loss (BPL) and adverse propagation properties which make them suitable only for short range communications. However, the frequency band of 28 GHz experiences less rain attenuation and absorption [6] compared with other mmWave frequency bands. This enables the 28 GHz band to be considered as a preferred band for cellular operators for future cellular communications. Additionally, the higher penetration and path loss can be coup with large antenna arrays with high directivity. This can be achieved due to the short wavelength of mmWaves which enables the utilization of large antenna arrays at both the transmitter and receiver side [6].

Generally, a single transmission point/cell serves a user in traditional wireless communications. However, in 3rd Generation Partnership Project (3GPP) release 10 the concept of coordinating multipoint transmission (CoMP) is introduced as a part of LTE-Advance technology, and that enables the user to receive signals from multiple transmission points. The coordination between different transmission points is based on the geographical location and the spatial separation. CoMP helps to improve the quality and the strength of the received signal [9-10]. Furthermore, the coordination can occur between the cells of the same base station or different base stations [9]. Coordinated Multipoint transmission can be categorized into following three categories, (1) Dynamic Point Selection (DPS) [11], (2) Joint Transmission (JT), and (3) Coordinated Beamforming/ Scheduling (CB/CS) [12]. A key drawback of CoMP is the huge quantity of data that is transferred between the coordinating transmission points.

2. Background Theory

2.1. Millimeter Waves

5G is expected to provide extremely high data rates for end users/devices. Researchers envision a thousand-fold increase in the data rate in 5G in comparison with the present technologies such as LTE-A. Higher data rates are can be achieved by increasing the bandwidth of the system. Traditional cellular communications utilize a part of the frequency band between 700 MHz and 4 GHz, that have limited bandwidth and is already occupied by the mobile operators.

The spectrum between 30 GHz and 300 GHz constitute the millimeter wave band. The wavelength of millimeter waves is in an order of few millimeters. The underutilization of the mmWave spectrum is due to the adverse conditions that affect the propagation [6]. Rain absorption, high penetration loss, low diffraction around the obstacles and a higher path loss are some of the key unfavorable factors that affect the propagation at millimeter wave frequencies. Additionally, in urban or dense urban city type environment, the communication links with only non-line-of-sight (NLOS) paths have a great challenge. As the signal attenuates more quickly at mmWave due to various adverse factors mentioned before in this article. Due to un-favorable propagation characteristics, the mmWave was envisioned as a solution for only short-range communications, particularly at the 60 GHz frequency [14].

Large unlicensed bandwidth (up to continuous 7 GHz) present at the 60 GHz band is of the primary reason for attraction, as this available bandwidth can be used for the future cellular communications system for providing enormous high data rates. The size of the bandwidth available at 60 GHz band is comparable to the bandwidth assigned for Ultra Wideband (UWB) purposes [15]. In addition the 60 GHz band is less restricted in terms of power limits. Additionally, 60 GHz frequency band utilizes greater power in comparison with Wireless Local Area Network (WLAN) and Wireless personal Area Network (WPAN). This is due to the 60 GHz frequency band regulation that allows the use of higher power. Therefore, the higher power can compensate for the greater path loss up to some extend at 60 GHz band [14]. This enables the 60 GHz frequency band to be a viable operating frequency for future wireless communications. However, there are other factors like high path loss, high rain attenuation and higher atmospheric absorption attributed to 60 GHz band; therefore, we have focused on 28 GHz band in this article.

In the sub-6 GHz frequency band, most of the spectrum is licensed and has given by the government for various purposes. It has resulted in a shortage of free available spectrum in sub-6 GHz band. Nevertheless, the un-utilized 28 GHz frequency band can help in alleviating the problem of capacity crunch by offering large free bandwidth [6]. However, there are some drawbacks of using mmWave frequency band. Based on the fundamental laws of physics, the frequency is inversely proportional to the pathloss and that results in the limited coverage at the 28 GHz frequency [16]. Absorption in the atmosphere due to the presence of rain, fog and mist also affects coverage at mmWave frequencies. Operation at mmWave frequencies and advancement in the semi-conductor technology has enabled to place a large number of antenna elements in a small cross-section, and it results in high directivity gain [6]. An antenna array with a large number of antennas can be utilized at both the transmitting and receiving sides due to the relatively small wavelength of mmWave. Additionally, electrically steerable antennas are also practical due to the development of CMOS RF circuits and that helps in improving the communication range [6].

2.2. Coordinated Multipoint (CoMP) Transmission

The 3GPP introduced coordinating multipoint technology as part of its Release 11 for LTE-Advanced systems [4]. The CoMP technology helps in establishing coordination among the multiple sectors of the same or different macro cellular base station. The coordinated communication is established by either selecting the sectors from the same macro site or from different macro sites. Intra-site coordination means two or more sectors/cells from the same macro site provide communication to the user. In inter-site CoMP, two or more sectors from different macro sites coordinate to provide the signal to the user. The coordinated transmission helps in alleviating the interference from adjoining cells and that helps in improving the spectral efficiency of the cell. The improvement in the quality of the signal (at the cell edge) occurs as the interfering signal coming from the other port is converted into a meaningful signal in the joint transmission technique. The cell-edge uses experience a poor quality of the signal as the received signal strength from multiple base stations are in a close range in terms of power. The utilization of the CoMP technology helps in improving the throughput, not only for the cell edge users only but also for the other users [17-18].

In traditional cellular macro-sites, a single sector/cell establishes a communication link to a particular user in the downlink direction. Here, a cell corresponds to a transmission point. However, LTE Release 11 aims to provide support for

coordinated transmission both on the uplink and downlink channel. CoMP overcomes the hindrance arising from the interference that is caused by neighboring cells [19]. CoMP can be classified into different categories. The categorization of CoMP depends on the presence of a backhaul connection between the transmission points that are coordinating. Scheduling also has a major impact on the CoMP subcategories [11]. The three sub-categories of the CoMP technology are 1) Dynamic Point Selection (DPS) [11], (2) Joint Transmission (JT) [19-20], and (3) Coordinated Beamforming or Coordinated Scheduling (CB/CS).

In case of DPS, a single transmission point from multiple coordinating TPs is selected during every transmission time interval. The user equipment reports the Channel State Information (CSI), the identity of the favorable transmission point and other essential information such as the resources required to the central node. In DPS, after the selection of the transmission point, the data is transmitted to UE only from the selected transmission point, and other coordinating transmission points stays in mute mode during the transmission [19]. Utilizing the joint transmission type CoMP technique helps in providing channel diversity. The simultaneous transmission of identical data from different TPs helps in achieving the channel diversity. However, a suitable combining technique must be utilized at the receiver to combine the data coming from different transmission points. The joint transmission technique can be either coherent or non-coherent [19]. The signals are precoded simultaneously in coherent transmission. However, in non-coherent transmission, the precoding of the signals are performed separately at different transmission ports. The user throughput can be improved by using a joint scheduler that utilizes the available resources in more efficiently in different load condition. However, the increase in the number of TPs increase the complexity of the joint transmission technique. In coordinated scheduling and coordinated beamforming, only the CSI is transmitted from various TPs. However, only one TP is used for the actual data transmission [20]. The interference can be eliminated by utilizing one of the CoMP techniques mentioned above. It is essential to strike a balance between the performance gain and the network upgradation cost [21]. As stated above, the increase in the number of coordinating transmission ports increases the complexity of the coordinating system. A major drawback of the CoMP technology is that there is a large amount of information exchange among the coordinating TPs. Consequently, CoMP operation requires higher channel estimation and strict synchronization among the coordinating TPs. The authors in [20] and [21] demonstrate that the spectral efficiency of the system is improved by utilizing the CoMP technique. CoMP transmission help in mitigating the interference in some cases, and provide spatial diversity.

2.3. Proposed Performance Metric

In the current Information and Communication Technology (ICT) industry there is a dire need to design energy or power efficient systems. Power efficiency can be achieved by having energy efficient architectures, energy efficient resource management, and by utilizing energy efficient radio technologies [22-23]. Utilizing simultaneous connections in addition to deploying small cells help in increasing the coverage of the network. In order to compute the consumption of power at a site, a power consumption model is presented at [24] and [25] and is given as:

$$P_{BS} = P_{const} + P_{load} * F, \quad (1)$$

In equation (7), P_{BS} is the total power consumption of the base station, P_{const} is the load-independent power consumption factor, P_{load} is the load-dependent power consumption factor and F is the load factor. For the BS operating at full load i.e. 100 %, the load factor F is equal to 1. The load-independent and load-dependent factors included in this study of this article are presented in Table .

Table 1. Parameters of power consumption model.

Parameter	Unit	Value
Transmit power per transmission port	dBm	46
Power consumption of DSP unit	W	100
Power Amplifier (PA) efficiency	%	45
Consumption of tranceiver's power	W	100
Consumption of rectifier's power	W	100
Consumption of fiber optic unit power	W	7.5

The relative power usage (P_{Rel}) in [%] is defined as follows:

$$P_{Rel} = \left(\frac{P_U - P_{Ref}}{P_{Ref}}\right) * 100\,, \quad (2)$$

In equation (8), P_U and P_{Ref} is the power usage of the case under consideration and power usage of the reference case, respectively. The power usage of the reference case (P_{ref}) and considered case (P_U) is computed as given in equation (9) and (10), respectively. In equation (9), T_{ref} is the time of the reference case and is defined as the time required for sending the reference amount of data (D_{Ref}) with a data rate of reference case (R_{Ref}) as given equation (11). Whereas, in equation (10), T_{Rel} is the relative amount of time needed to transmit the reference data (D_{Ref}) and is defined as a ratio of T_U and T_{Ref} as expressed in equation (12), where T_U is the time of the considered case and is defined in equation (13) as the time required for sending the reference amount of data (D_{Ref}) with a data rate of case under consideration (R_U).

$$P_{ref} = P_{BS} T_{ref} \quad (3)$$

$$P_U = P_{BS} T_{Rel} \quad (4)$$

$$T_{Ref} = D_{Ref}/R_{Ref} \quad (5)$$

$$T_{Rel} = T_U/T_{Ref} \quad (6)$$

$$T_U = D_{Ref}/R_U \quad (7)$$

Details of the other performance metric i.e. area spectral efficiency, and relative capacity gain can be found at [1].

3. Simulation Environment, Cases and Parameters

3.1. Simulation Environment

A three-dimensional ray tracing simulator called "sAGA" is developed by the authors using a MATLAB platform for the simulation purpose. An 'Image Theory' based ray tracing algorithm is to find the propagation paths between the transmitter and the receiver. A tracing simulator requires 3D building data to accurately determine the paths between the transmitter and the receiver. For this study work, a Krunuuhaka region from Helsinki is considered. The two-dimensional map of considered area is shown in Figure 1. This type of environment illustrates an urban

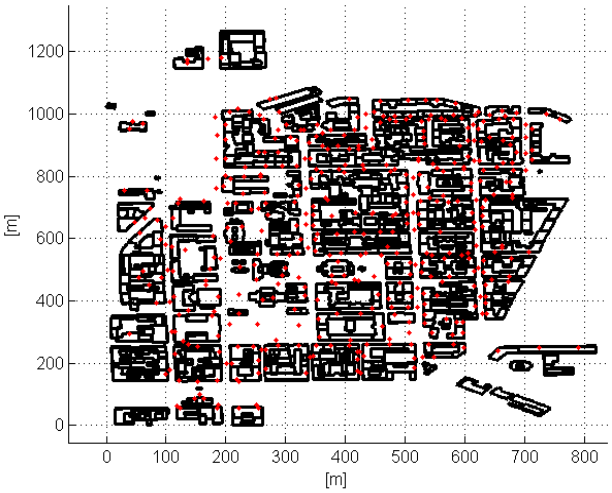


Figure 1. Simulation area.

macro cellular environment. It is assumed that users are located both in indoor and outdoor locations. For indoor users, it is assumed that there are no walls inside the indoor environment. However, BPL and ceiling penetration loss are considered for indoor users. For the outer wall with the 40 cm thickness, the wall penetration loss is 26.5 dB at 28 GHz frequency [13]. As a fairly large number of miniature antenna elements can be placed inside a small area at 28 GHz, therefore, an additional gain of 16 dB is assumed in our analysis. The basis of this additional gain is the assumption that a large array of an antenna can provide additional diversity gain, beamforming gain and combining gain.

There are ten base stations with three sectors. The height of the antenna is set at 30 m for all cells. The illustration of the antenna locations of the antennas is given in Figure 3. The users (850) were placed in the environment with approximately 85% of the users located indoors and the remaining 15% located outdoors. The indoor users are located on different floors. A floor height of 3 meters was considered for the indoor environment. The locations of users on different levels of the building are elucidated in Table.

Table 2 . User distribution.

	Number	Percentage
Total RX points	850	100
Outdoor	128	15.06
Indoor	722	84.94
Ground floor	269	31.65
Second floor	259	30.47
Fifth floor	160	18.82
Seventh floor	34	4

A. Simulation cases

Cellular networks are generally deployed with the macro layer to provide the coverage and the basic capacity to the network, and such networks are called homogeneous networks. Therefore, for the research work of this article, only homogeneous scenarios with only macro sites are considered. The simulation cases considered for this work are presented below.

3-sector site (Reference case): The reference case illustrates a case with the macro site, where only one cell (acting as the serving cell) provides connection to the user. A serving cell is selected on the basis of the strength of the received signal. In order to restrict the interference from the neighboring cells, a downtilt of 9° was considered for all the sectors, as the sites are located fairly close to each other

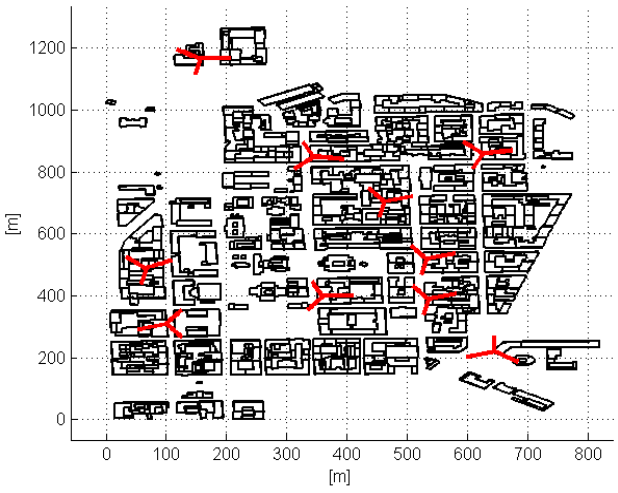


Figure 2. Three-sector site plan.

Intra-Node 2 TP: The same macro base station provides coordinated transmission by virtue of utilizing two Transmission Points (TPs). In this context, a TP represents a single sector/cell. The utilization of the same macro cellular site eliminates the requirement for a backhaul link between the coordinating cells.

Intra-Node 3 TP: In this case there is a coordinated transmission from the 3 cells of the same macro site.

Inter-Node 2 TP: This case represents a scenario where the coordinated transmission is provided from two sectors from different macro sites. As a result, a backhaul link is necessary in such a scenario.

Inter-Node 3 TP: This scenario illustrates the coordinated transmission from three cells belongs to different sites.

Inter-Node 4 TP: It involves transmission from four cells which may belong to different sites.

The antenna radiation pattern is obtained using an extended version of the 3GPP antenna model. The key parameters utilized for the antenna modeling are summarized in Table 3.

Table 3: Antenna model parameters

$HPBW_H$	$HPBW_V$	FBR_H	SLL_V	A_M
65°	6.2°	30 dB	-18 dB	18.23 dBi

Other simulation parameters are provided in Table 4.

Table 4. General simulation parameters

Parameters	Unit	Value
Frequency	GHz	28
System bandwidth	MHz	20
TX power	dBm	46
Number of cells	No.	30
Antenna height	m	30
Building penetration loss	dB	26.5
UE noise figure	dB	8

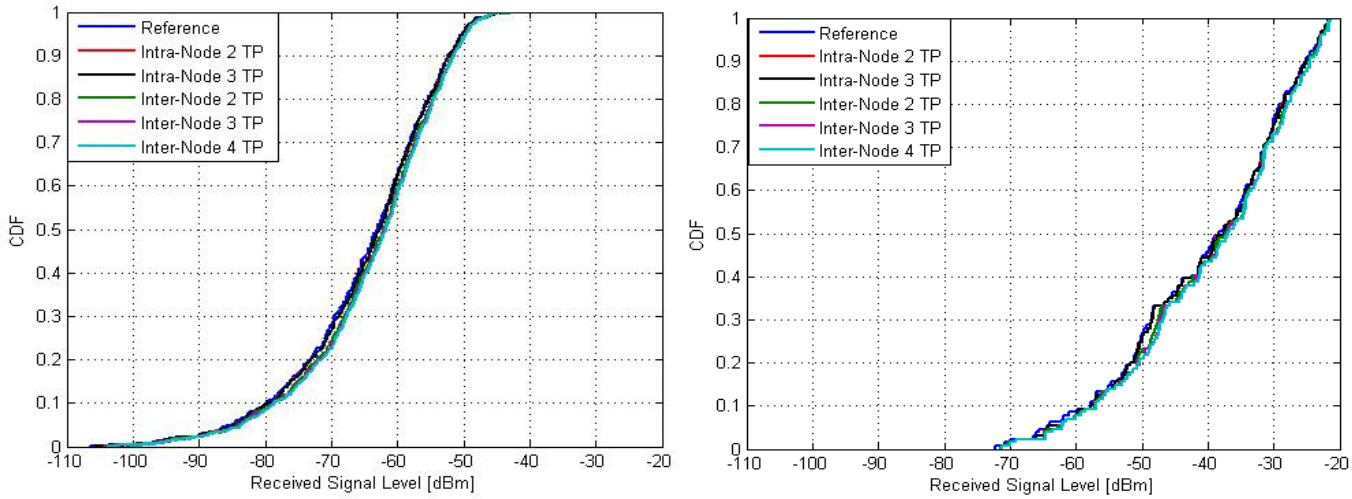


Figure 3. Received signal level, (a) CDF plot of indoor users, (b) CDF plot of outdoor users.

4. Simulation Results and Discussions

The first metric considered for the analysis is the received signal strength. Indoor and outdoor users are separately analyzed in this paper. Figure 3 (a) shows the CDF plots of received signal level in [dBm] for different considered cases for indoor users. It can be seen in Figure 3 (a) that almost identical received signal levels were obtained by reference cases and other coordinated multipoint transmission cases. Mean received level of -64.85 dBm is obtained with reference case, and with marginal improvement the received signal strength of -64.57 dBm and -64.56 dBm is achieved with two and three points intra-node CoMP transmission, respectively. Similarly, the maximum mean received signal strength of -63.59 dBm is acquired with four points inter-node CoMP case, and that is again not significantly better compared with the reference case. The Figure 3 (b) shows the CDF plots of received signal level in [dBm] for outdoor users. It is clearly evident that outdoor users have stronger signal strength compared with indoor users, as the indoor signal experiences the building penetration loss. Mean RX level of -40.46 dBm is obtained with the reference case for outdoor users, and mean received signal strength of -39.37 dBm is acquired with four points inter-node CoMP case.

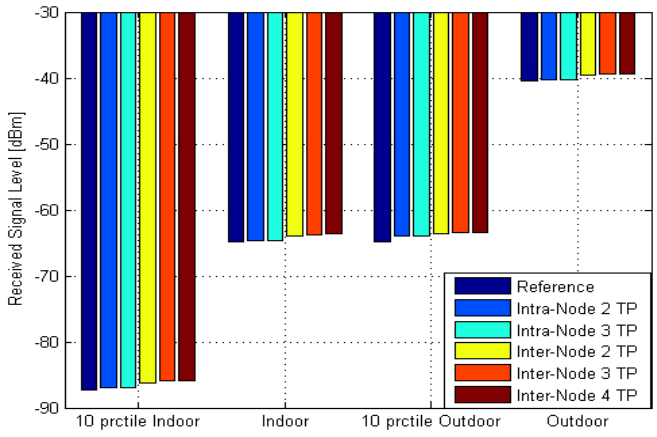
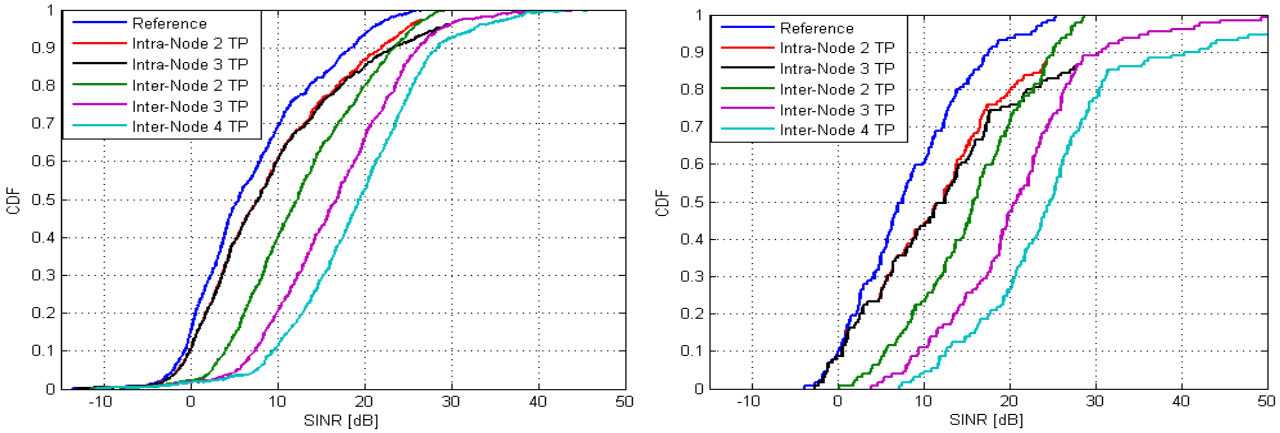


Figure 4: Bar plot of mean received signal levels.

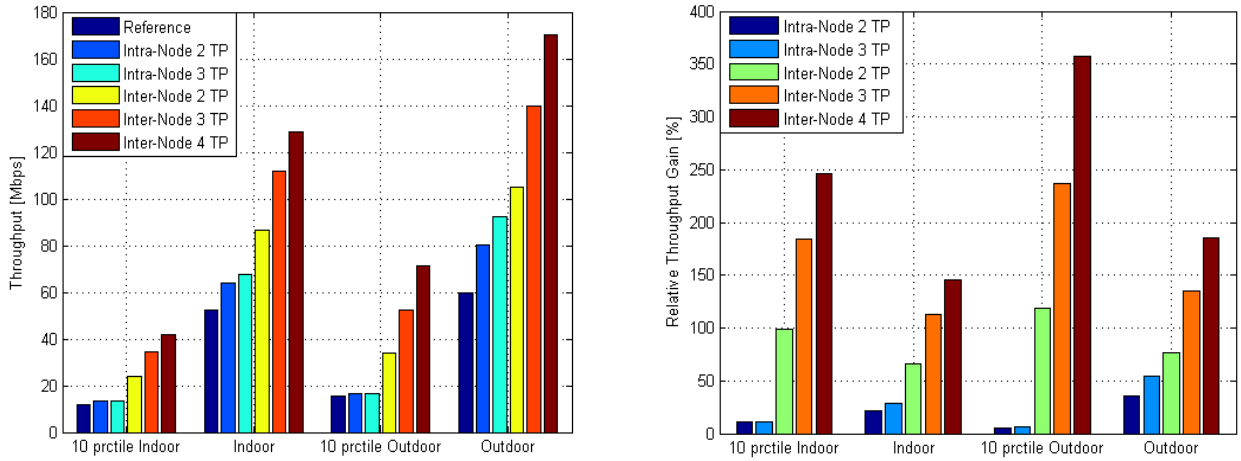
The performance of the users within the 10th of the sample data are more critical and crucial. Therefore, we have separately targeted the cell edge users (10th percentile data) for our later analysis and comparison of performance. Therefore, we have separately targeted the cell edge users (10th percentile data) for our later analysis and comparison of performance. Figure shows the bar plot of mean received signal level for indoor and outdoor users considering the 10th percentile users and the whole data set. It is interesting to see that mean received signal of indoor users considering the whole data set is almost identical to the signal levels of outdoor users within the 10th percentile of data. Only received signal level does not tell about the actual user experience, and in order to have more deep understanding of user experience, a quality metric i.e. SINR needs to be analyzed along with signal strength.

Figure 5 (a) and figure 5 (b) shows the CDF plots of the SINR of indoor users and outdoor users for different considered cases, respectively. The reference case considered in this article offers a mean SINR of about 7.1 dB and 8.37 dB in the targeted area for indoor and outdoor users, respectively. It can be seen in figures 5 (a) and 5 (b) that the CoMP functionality has improved the SINR compared with the reference case, and the inter-node CoMP deployment is found more beneficial compared with intra-node CoMP performance. In case of intra-node CoMP deployment, it was found that increasing the number of joint transmission points (cells) from two to three shows no significant improvement in SINR, which is due to large spatial separation between the sectors of the same site. For indoor users, the mean SINR of about 9.12 dB and 9.7 dB is achieved with two and three points intra-node CoMP functionality, respectively. However, in case of inter-node CoMP functionality the gain in improving the received SINR enhances with the increase in number of joint transmission points. For indoor users, the mean SINR of 12.83 dB, 16.72 dB, and 19.30 dB is acquired from 2-, 3-, and 4-point joint inter-node CoMP transmission, respectively. A higher mean SINR values are achieved in outdoor environment, and similarly the CoMP technique improves the SINR of outdoor users as well. A mean SINR of 25.6 dB attained with 4-point joint inter-node CoMP transmission for outdoor users.



(a) (b)

Figure 5. Signal to interference plus noise ratio, (a) CDF plot of indoor users, (b) CDF plot of outdoor users.



(a) (b)

Figure 6. Throughput, (a) Mean user throughput, and (b) Relative throughput gain.

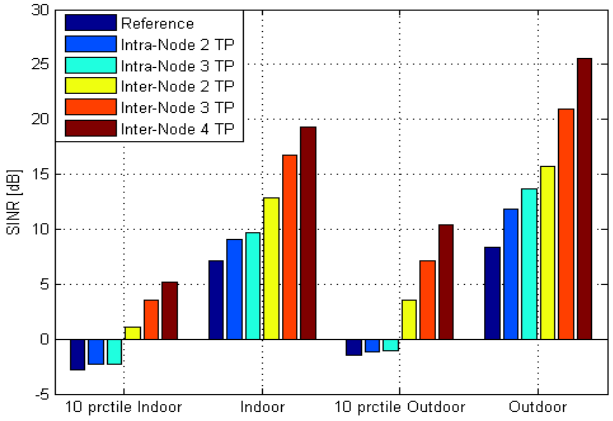


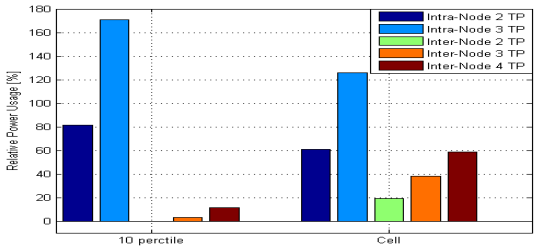
Figure 7. Bar plot of SINR.

As stated earlier, that the users within 10th percentile data are vital and figure (5) shows the bar plot of mean SINR for indoor and outdoor users considering the 10th percentile users and the whole data set. Considering the SINR results presented in 5 (a) and 5 (b) it can be said that the CoMP deployment has improved the SINR of the cell edge users by considerable margin. The mean SINR of indoor cell edge users is improved from -2.84 dB to -2.28 dB, and that means a gain of just 0.56 dB. Again, no additional sign of improvement is witnessed by increasing a number of TPs from two to three for intra-node CoMP transmission. However, the inter-node CoMP deployment improves the indoor user experience in downlink direction and provides a gain of 3.96, 6.42, and 7.98 dB by 2, 3, and 4 points inter-node joint transmission, respectively. Considering only 10th data, even higher gains are achieved in an outdoor environment and a gain of 4.99, 8.6, and 11.8 dB is provided by 2, 3, and 4 points inter-node joint transmission, respectively compared with the reference case. It shows the potential of CoMP deployment in improving the user quality at the cell edge.

The Figure 6 (a) shows the mean throughput of the users for different considered cases for indoor and outdoor environment, whereas the Figure 6 (b) shows the relative throughput gain of different CoMP cases with respect to the reference case. The "10th prctile" in Figure depicts the throughput of the users in the lower 10 percentile users' throughput distribution. In 6 (a), although the increase in the absolute value of throughput (capacity) by CoMP is small at cell edge level, however the relative capacity gain coming from CoMP functionality is higher for cell edge users compared with overall cell users as shown in 6 (b). It is also learned from the results presented in Figure that the intra-node CoMP deployment is less efficient for cell edge users, and does not provide any significant gain i.e. around 11.01 – 11.4 % for indoor users, and 5.8 – 6.1 % for outdoor users, respectively. However, inter-node CoMP deployment can offer the maximum relative capacity gain of about 98.2 %, 183.7 %, and 246.6 % with

2, 3, and 4 points joint scheduling and joint transmission for indoor users. Similarly, for outdoor users in the lower 10 percentiles, CoMP offered relative capacity gain of about 118.6 %, 237 %, and 357 % with 2, 3, and 4 points inter-node CoMP transmission.Table shows the results of the overlapping zone with 3 dB, 5 dB, and 7 dB window for different considered cases. The information about the number of available servers (carriers) within the window of certain dBs is commonly used by mobile operators to evaluate the design and performance of the network, and is also used as a metric of network quality during the planning process. The percentage of the area with the single server within the window of certain dBs shows the clear dominance area. However, an overlapping between the cells is required to support the handover between them. An area overlapped by multiple servers is affected by higher interference, and is called as "polluted" area with multiple servers; therefore the SINR is directly related to carrier overlapping. In Table it can be seen that the cell dominance i.e. single server area improves with CoMP functionality. With 5 dB window, the reference case has 67.48 % of clear cell dominance area which is then extended to 75.64 % with intra-node 2 point transmission. It can be also seen that no additional improvement is brought to cell dominance area with intra-node 3-point transmission, and therefore there was no additional improvement in SINR results with intra-node 3-point transmission, which were earlier presented in figure (7). However, for the cases of inter-node CoMP deployment, the cell dominance area continued to improve with the additional number of coordinated points. For the case of 5 dB window the inter-node 3 point CoMP transmission provides cell dominance area of 97.55 %, which is then further improved to 100 % with inter-node 4 point CoMP transmission. Similarly, the statistics of the overlapping zone with 3 dB and 7 dB for different cases can be found at Table . Interestingly, the results presented in Table show that on one hand CoMP extends the cell dominance area, on the other hand it diminishes the handover area with small difference of levels between the available servers, and it will help in avoiding the ping-pong effect during the handover procedure.

Figure shows the relative power usage of different considered CoMP cases with respect to the reference case using 10th percentile users' data and all users' data set. In case of a reference case, a user is connected with only one transmission point, whereas in CoMP the user is connected with multiple transmission ports, therefore utilizes the radio and power resources of multiple sources. The relative power usage of 0 % means a same power is required in the considered case as used in the reference case, and a 100 % relative power usage means twice the power is required in case under consideration with respect to the reference case.



It is already established from the SINR and throughput results presented in figure (8) and Figure that the CoMP transmission improves the SINR of the user which in turn improves the user throughput. However, the Figure shows that intra-node CoMP transmission is not power efficient solution. It is learned from relative power usage results presented in Figure that while using two point and three point intra-node CoMP transmission, 81.2 % and 170.1 % relative additional power is required in comparison with the reference case considering cell edge users. Similarly, 61 % and 125.5 % relative additional power is required in comparison with the reference case considering whole cell data. On the other hand, inter-node CoMP techniques are quite much power efficient and work better for serving cell edge users. Two-point inter-node CoMP transmission is even more power efficient compared with reference case, and three point and 4-point inter-node CoMP has a relative power usage of only 3.2 % and 11.2 %, respectively for 10th percentile data. The summary of the results considering the whole data set is given in Table 6.

5. Conclusion

It is learned from the results presented in this article that the coordinated multipoint transmission and joint scheduling technique help in alleviating the problem of bad coverage and poor signal quality for the users located at the cell edge. It was found that CoMP is more beneficial for the cell edge users in comparison with the users located near the base station. The conventional case of 3-sector site was used as a reference case to compute the relative throughput gain for different CoMP configurations. The findings of this article revealed that utilizing two TPs in CoMP mode from the same base station provides the relative throughput of 10.32 % for the cell edge users. However, significant gain is not witnessed by utilizing three cells from the same base over two TPs from the same base station. Utilization of two, three and four TPs from different coordinating base stations yields a mean relative throughput gain of 100.5 %, 190.5 % and

Table 5. Results of overlapping zone.

Case	**Single server 3dB (%)**	**Two servers 3dB (%)**	**≥Three servers 3dB (%)**	**Single server 5dB (%)**	**Two servers 5dB (%)**	**≥Three servers 5dB (%)**	**Single server 7dB (%)**	**Two servers 7dB (%)**	**≥Three servers 7dB (%)**
Reference	80.65	16.32	3.03	67.48	23.43	9.09	55.59	29.84	14.57
Intra-Node 2 TP	85.78	12.47	1.75	75.64	18.18	6.18	65.04	24.13	10.84
Intra-Node 3 TP	85.78	12.47	1.75	75.64	18.18	6.18	65.04	24.13	10.84
Inter-Node 2 TP	100.0	0	0	97.55	2.21	0.23	90.68	7.69	1.63
Inter-Node 3 TP	100.0	0	0	100.0	0	0	99.53	0.47	0
Inter-Node 4 TP	100.0	0	0	100.0	0	0	100.0	0	0

Table 6 Results summary

Case	Mean RX Level [dBm]	Mean SINR [dB]	Relative SINR Gain [dB]	Mean Cell Spectral Eff. [bps/Hz]	Mean Throughput [Mbps]	Relative Throughput Gain [%]	Relative Power Usage [%]
Reference	-61.19	7.29	0	2.66	53.38	0	0
2Point Intra	-60.91	9.53	2.24	3.31	66.34	24.27	60.94
3Point Intra	-60.90	10.30	3.01	3.55	71.00	33.0	125.57
2Point Inter	-60.29	13.26	5.97	4.47	89.40	67.47	19.43
3Point Inter	-60.04	17.36	10.07	5.79	115.85	117.0	38.25
4Point Inter	-59.95	20.24	12.95	6.73	134.76	152.42	58.46

260 %, respectively, for the cell edge users. Similarly, over the whole cell area the coordinated multipoint transmission from 2, 3, and 4 TPs from different nodes can offer the mean relative throughput gain of 67.5 %, 117 %, and 152.5 %, respectively.

It was also found that utilizing two TPs and three TPs intra-node CoMP transmission requires 81.2 % and 170.1 % relative additional power in comparison with the traditional single point transmission considering only cell edge users. Whereas the inter-node CoMP transmission is found much more power efficient compared with single point transmission and intra-node Three TPs and four TPs inter-node CoMP has additional 3.2 % and 11.2 % power usage in comparison with single point transmission.

It was also revealed through simulation results that CoMP deployment clearly improves the cell dominance area, and shrinks the cell border area which is overlapped by multiple servers. The results presented in this paper show the potential of CoMP transmission in overcoming the problem of cell edge users. It is learned that intra-node CoMP solution is not a power efficient technique. However, inter-node CoMP does not only improves the user experience i.e. SINR and throughput; rather it is also power efficient to use inter-node coordinated multipoint transmission. The simulation results given in this research article have been collected for an urban area of Helsinki city with the certain simulation parameters and assumptions. Therefore, performance results can vary depending upon simulation environment, setup, modeling errors and impairment, and system parameters.

References

[1] M. U. Sheikh, R. Biswas and J. Lempiaeinen, "Performance Evaluation of Coordinated Multipoint Transmission at 28 GHz Frequency Using 3D Ray Tracing," 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018, pp. 1-6.

[2] Cisco white paper, Visual Networking Index, Feb. 2016.

[3] M. U. Sheikh, J. Lempiäinen and H. Ahnlund, "Advanced Antenna Techniques and Higher Order Sectorization with Novel Network Tessellation for Enhancing Macro Cell Capacity in DC-HSDPA Network", International Journal of Wireless & Mobile Networks. vol. 5, no.5, pp. 65–84.

[4] 3GPP, TS 36.300 "E-UTRA and E-UTRAN Overall description: Stage 2" (Release 11), V11.5.0, April. 2013.

[5] H. Zhao , R. Mayzus , S. Sun , M. Samimi , J. K. Schulz , Y. Azar , K. Wang , G.N. Wong , F. Gutierrez Jr. and S.T. Rappaport , "28 GHz Millimeter Wave Cellular Communication Measurements for Reflection and Penetration Loss in and Around Buildings in New York City", Proc. IEEE Int. Conf. Commun. , pp.1 -6 , 2013.

[6] T. S. Rappaport , J. N. Murdock and F. Gutierrez , "State of the art in 60 GHz Integrated Circuits & Systems for Wireless Communications" , Proc. IEEE , vol. 99 , pp.1390 -1436 , 2011.

[7] Z. Pi, and F. Khan, "An introduction to Millileter-wave Mobile Broadband Systems", IEEE Communications Magazine, vol. 49, no. 6, pp. 101-107, June 2011.

[8] S. K. Yong and C.-C. Chong, "An overview of Multigigabit wireless through millimetre wave technology: potentials and technical challenges," EURASIP journal on wireless communications and networking, 2007.

[9] T. Abe, Y. Kishiyama, Y. Kakura, and D. Imamura, "Radio Interface Technologies for Cooperative Transmission in 3GPP LTE-Advanced", IEICE Transaction on Communications., vol. E94-B, no. 12, pp. 3202-3210, Dec. 2011.

[10] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," IEEE Communications Magazine, vol.50, no.2, pp.148-155, February 2012.

[11] 3GPP TR36.814 V2.0.1 : "Evolved Universal Terrestrial Radio Access (E-UTRA): Further Advancements for E-UTRA Physical Layer Aspects", March 2010.

[12] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," IEEE Communications Magazine, vol.49, no.2, pp.102-111, February 2011.

[13] B. G. Choi, W. H. Jeong, K. S. Kim, "Characteristics Analysis of Reflection and Transmission According to Building Materials in the MillimeterWave Band", Recent Advances on Electroscience and Computers, pp. 154-158, 2015.

[14] S. K. Yong and C.-C. Chong, "An overview of Multigigabit wireless through millimetre wave technology: potentials and technical challenges," EURASIP journal on wireless communications and networking, 2007.

[15] ETSI, Accessed on 21/03/2016: http://www.etsi.org/technologies-clusters/technologies/radio/ultra-wide-band

[16] Digi International, Accessed on 21/03/2016: http://www.digi.com/technology/rf- articles/rf-basics.

[17] 3GPP, TR 36.819 "Coordinated Multi-Point Operation for LTE Physical Layer Aspects."

[18] K. M. Karakayli, G. J. Foschini, and R. A. Valenzuela, "Network Coordination for Spectrally Efficient Communications in Cellular Systems," IEEE Wireless Communication, vol. 13, no. 4, pp. 56–61, Aug 2006.

[19] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," IEEE Communications Magazine, vol.50, no.11, pp.44-50, November 2012.

[20] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," IEEE Communications Magazine, vol.50, no.2, pp.148-155, February 2012.

[21] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," IEEE Communications Magazine, vol.49, no.2, pp.102-111, February 2011.

[22] Su, X. , Sun, E. , Li, M. , Yu, F. and Zhang, Y. (2013) A Survey on Energy Efficiency in Cellular Networks. Communications and Network, 5, 654-660. doi: 10.4236/cn.2013.53B2117.

[23] O. Adigun, and C. Politis, "Energy Efficiency Analysis for MIMO Transmission Schemes in LTE," Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2011 IEEE 16th International Workshop on, Kyoto, 10-11 June 2011, pp. 77-81.

[24] S.F. Yunas, J. Niemela, M. Valkama, T. Isotalo, "Techno-economical analysis and comparison of legacy and ultra-dense small cell networks," in Local Computer Networks Workshops (LCN Workshops), 2014 IEEE 39th Conference on , vol., no., pp.768-776, 8-11 Sept. 2014.

[25] M. U. Sheikh, J. Sae, and J. Lempiäinen, "In Preparation Towards Future Cellular Network: The Detailed Analysis of Macro and Micro Site Densifcation and Sector Densification", Journal of Telecommunication Systems. vol. 65, no.4, pp. 621–636, August 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 125-140 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Conducted and Radiated Interference on Interconnection's Lines

Patricio E. Munhoz-Rojas*

Department of Research, Development and Innovation, Lactec, Brazil

ARTICLE INFO

Article history:
Received: 28 November, 2018
Accepted: 12 January, 2019
Online : 29 January, 2019

Keywords:
Electromagnetic potentials
Electromagnetic interference
Electromagnetic disturbances

ABSTRACT

There exists a profound difficulty of communication between the people that works in the EMC area in circuit terms and the people that works in field terms.
In this paper we show that when the matter is predominantly distributed along a certain direction in space, as for transmission lines, the electromagnetic field can be divided in two modes, each of them with two degrees of freedom, that are practically independent: a longitudinal TM (transverse magnetic) mode and a transversal TE (transverse electric) mode. We also show that two degrees of freedom of the longitudinal mode are the ones that are described by circuit's theory.
This formulation is based on the observation that, when the matter is macroscopically described by constitutive laws, the electromagnetic field within the matter can be fully characterized in terms of the potential fields, in total four degrees of freedom. Using the above formulation, we put forward a generalized formulation of the coupling of an external electromagnetic field to a transmission line, valid in any time scale.
We apply the above concepts to study, in a common theoretical framework, the iconic case of the conducted and radiated interferences on a transmission line, and we show that:
1-Differently than what is normally assumed in standard transmission-line theory, the normal operation mode and the internally-produced electromagnetic field are predominantly a longitudinal TM mode;
2-The longitudinal mode is affected by both the conducted disturbances and the radiated disturbances; while
3- The transverse mode is affected only by the radiated disturbances.
Then, only for systems where the longitudinal mode is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.
Also, to further illustrate the interpretation power of this formulation, we present some other application examples.

1. Introduction

This paper is an extension of a work originally presented in the 2018 Joint IEEE EMC & APEMC, which took place in Singapore from 14 to 17 May 2018 [1], and its main purpose is to present in a more detailed form both the general theoretical framework and the demonstration that the circuit's theory applies to the two degrees of freedom of the so-called longitudinal mode: the scalar potential "Φ" in the conductor and the magnetic potential "A_{mz}" along the conductor (the current "i" in the conductor is related to A_{mz} through the concept of inductance).

In reference [1], this theory was applied to study conducted and radiated interferences in transmission lines, because there exists confusion among the people that work in the area, in different time scales; which is caused, as mentioned there, by a rather loose definition of conducted and radiated disturbances.

But nor the answer to the question, why do we feel that they are rather loosely defined? was fully explained in the text of reference [1], neither the primary cause of the confusion was identified.

In reference [1] it was recalled that "in IEC, the disturbing electromagnetic fields are divided into: *conducted disturbances* (IEV 161-03-27) and *radiated disturbances* (IEV 161-03-28).

*Patricio E. Munhoz-Rojas, Email: patricio@lactec.org.br

The essential difference between these disturbances, as established in the above definitions, resides in the manner how the energy is transferred to the conductors:

- For conducted disturbances, IEV 161-03-27 says that the energy is transferred via one or more conductors;

- For radiated disturbances, IEV 161-03-28 says that the energy is transferred through space in the form of electromagnetic waves. However, it notes that "The term "radiated disturbance" is sometimes used to cover induction phenomena"."

We feel that the above definitions of conducted and radiated disturbances are rather loosely defined because they appeal to an intuitive concept of energy on the part of the reader, taking advantage of the fact that the electromagnetic energy is very seldom calculated in interference problems. Besides that, the concept of electromagnetic energy of the people that works in circuit terms is very different than the concept of the people that works in field terms.

IEV definition of electromagnetic energy 121-11-64 says that "The energy associated with an electromagnetic field, in a linear medium within a domain V, is given by the volume integral

$$\frac{1}{2}\int_V \left(\vec{E}\cdot\vec{D}+\vec{H}\cdot\vec{B}\right)dV$$

where **E**, **D**, **H** and **B** are the four vector quantities determining the electromagnetic field".

This definition in field terms is not easy to understand for the people working in the low-frequency regime, which is used to work in circuit terms. Also, it is not frequently known what the relation between the above defined concept of electromagnetic energy and the concept of energy and power, used in circuit terms, is.

Then, we can trace back the primary cause of the confusion to the fact that some people work in circuit terms, while other people work in field terms.

As mentioned in reference [1], this is the origin of the different approach and the rather different language used by the people working in the low-frequency regime, as the power quality area, and the people working in the high-frequency regime, both in the lightning protection area and in the EMC area, facts that cause a certain degree of confusion and make it difficult the communication among the people working in these different areas.

Also, as very well observed in reference [2], "Within IEC, power quality is treated within the standards on electromagnetic compatibility (EMC)". This is because EMI problems, both in the low-frequency regime and in the high-frequency regime, are quantified and measured in terms of *currents* in the conductors and *voltages* between conductors, which is partly due to the widely accepted belief that the EMI on transmission lines is a completely known problem, and it can be simulated using various circuit simulation software's.

But this belief assumes that the problem of determining to what kind of electromagnetic systems the circuit theory is applicable, or which are the limits of the validity of the circuit theory, is a solved problem. Assumption that is by no means valid [3].

In this paper, as in reference [1], we first demonstrate that the characterization of the electromagnetic field in terms of the *electric scalar potential* "Φ" and of the *magnetic vector potential* "$\mathbf{A}_m$" [4], which Maxwell called *electromagnetic momentum* [5], when the matter can be macroscopically described by *constitutive laws*, is totally general. This characterization has in total four degrees of freedom.

Second, we demonstrate that the circuit's theory applies to the two degrees of freedom of the so-called longitudinal mode: the scalar potential "Φ" in the conductor and the magnetic potential "A_{mz}" along the conductor (the current "i" in the conductor is related to A_{mz} through the concept of inductance).

This paper is organized as follows:

In part 2, which closely follows the description given in part II of reference [1], we first review the different forms of description of the electromagnetic field within the matter [4, 6-9] showing that all of them have in common the equations involving the vector fields **E** (electric field strength) and $\mathbf{B}_m$ (magnetic flux density), which are commonly called "*electric field*" and "*magnetic field*".

Also, as in reference [1], it is shown that:

- When the matter can be macroscopically described by *constitutive laws*, which are relations between the other fields (**H**, **D** and $\mathbf{J}_{efr}$) needed to describe the electromagnetic field within the matter and the fields **E** and $\mathbf{B}_m$; then, the electromagnetic field within the matter is fully characterized by the fields **E** and $\mathbf{B}_m$ (six degrees of freedom); and

- As the common equations for the fields **E** and $\mathbf{B}_m$ can be solved in terms of the so-called *magnetic vector potential* "$\mathbf{A}_m$" and *electric scalar potential* "Φ", in the case when the matter is described by constitutive laws, the number of degrees of freedom needed for the characterization of the electromagnetic field within the matter can be reduced from six (**E**, $\mathbf{B}_m$) to four ($\mathbf{A}_m$, Φ).

Then, as it happens in the vacuum [10], when the matter is macroscopically described by *constitutive laws*, the electromagnetic field within the matter is fully described by these two potential fields $\mathbf{A}_m$ and Φ.

The extension made in this paper mainly refers to alert the reader on the differences, arising from the different formulations of electromagnetic theory and not always acknowledged, between the formulations utilized by the different application software's.

In part 3, which closely follows the line of reasoning of part III of reference [1], we first show that when the matter is predominantly distributed along a certain direction in space, as for transmission lines, the four degrees of freedom of the potentials can be separated into two independent modes, each of them with two degrees of freedom:

- a *longitudinal mode* constituted by Φ and the component of $\mathbf{A}_m$ along the line, which is a TM (transverse magnetic) mode [11], and

- a *transversal mode* constituted by the two components of $\mathbf{A}_m$ transversal to the line, which is a TE (transverse electric) mode [11].

As in reference [1], it is also shown that the longitudinal mode is the one described by circuit theory and the relation between its two degrees of freedom is given by Kirchhoff's laws, which are deeply related to the existence and predominance of the longitudinal mode.

As an extension made in this paper, using the longitudinal TM mode as the fundamental building block, instead of the TEM (transverse electromagnetic) mode as in traditional transmission-line theory [12], and assuming that the longitudinal and the transversal modes are practically independent, we present the derivation of a generalized theory of the electromagnetic field coupling to a multiconductor line, in time domain, that, as usual, predicts the propagation of the scalar potential and the current along the line.

For the longitudinal TM mode, as in traditional multi-conductor transmission-line theory [12-14], the line voltages have a unique value, independent of the integration path from the reference to the conductor.

The independent transversal mode produces additional induced voltages (integration path dependent) between the conductors of the line.

As mentioned in reference [1], this treatment can also be extended to lines with imperfections, discontinuities or with conductors attached perpendicular to the line, such as, the terminations, equipment connections or groundings [15].

Also, as an extension and an application of the concept of longitudinal mode, the derivation of a general theory of the coupling of an external electromagnetic field to a conductor line is presented. This theory can be applied to different problems, such as, electromagnetic neural stimulation [16-20] and the calculation of lightning-induced voltages [13]. In order to not deviate the attention away from the main purpose of the paper and to avoid too many mathematical derivations in the main body of the paper, we present it in Annex 1.

The generalized theory of the electromagnetic field coupling to a multiconductor line, under the proper simplifications, reduces to the standard coupling theories [12-14,21]. In Annex 2, we present a detailed comparison of this generalized theory with the most important classical theories of the electromagnetic field coupling to a multiconductor line.

In part 4, as in part IV of reference [1], we apply the above theory to analyze the interference on a transmission line produced by external disturbances, which are commonly classified into conducted and radiated disturbances.

As in reference [1]:

- We assume that the *normal operation mode*, which is driven by normal lumped external excitation sources, is a longitudinal mode;

- We divide the externally produced disturbances in two classes: *longitudinal mode disturbances* and *transversal mode disturbances*;

- We divide the longitudinal mode disturbances in two classes: the *scattered* and the externally produced; these last ones are also divided in two classes: the remotely produced and that produced by the *impressed current*, which is injected by lumped external sources.

We show that:

- the *conducted disturbances* are longitudinal mode disturbances that affect only the longitudinal mode; but,

- the *radiated disturbances* are composed of longitudinal mode disturbances and transversal mode disturbances, both of which affect the longitudinal mode; while the transversal mode is only affected by the transversal mode disturbances.

This is the reason why, only when the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.

Also, this explains why current injection and capacitive clamp testing methods represent only the effect of disturbances, both conducted and radiated, on the longitudinal mode.

In part 5, in order to illustrate the interpretation power of this approach, we present the results of other application cases together the important practical and engineering conclusions that has gone unnoticed in other calculations made with previously proposed approaches/software tools. Finally, in part 6 we present our summary and our main conclusions.

2. Description of the Electromagnetic Field

The description of the electromagnetic field given in this paper closely parallels the description given in reference [1] having only being added some complementary explanations, being (1) to (8) the same of reference [1] and our (13) is equal to (9) of reference [1].

The Maxwell-Hertz classical formulation of the electromagnetic theory [4,6], which refers to 5 vector fields and 1 scalar field (**E**, **H**, **D**, $\mathbf{B}_m$, $\mathbf{J}_{efr}$ and ρ_{efr}), is commonly called *Maxwell equations* (IEV 121-11-62):

$$\nabla \cdot \vec{B}_m = 0; \qquad \nabla \cdot \vec{D} = \rho_{efr};$$
$$\nabla \times \vec{E} = -\frac{\partial \vec{B}_m}{\partial t}; \qquad \nabla \times \vec{H} = \vec{J}_{efr} + \frac{\partial \vec{D}}{\partial t}. \tag{1}$$

Where the four quantities determining the *electromagnetic field* (IEV 121-11-61) are: **E** the *electric field strength* (IEV 121-11-18), **H** the *magnetic field strength* (IEV 121-11-56), $\mathbf{B}_m$ the *magnetic flux density* (IEV 121-11-19) and **D** the *displacement* (IEV 121-11-40). The vector field $\mathbf{J}_{efr}$ is the *electric (conduction) current density* (IEV 121-11-11) and the scalar field ρ_{efr} is the *electric charge density or volumic (electric) charge* (IEV 121-11-07), which according to IEV 121-11-61 are needed to characterize **the electric and magnetic conditions of a material medium** together the *electromagnetic field*.

Using the following definitions:

$$\rho_{efr} \equiv -\nabla \cdot \vec{M}_{efr};$$
$$\vec{J}_{efr} \equiv \frac{\partial \vec{M}_{efr}}{\partial t} + \nabla \times \frac{\vec{M}_{mfr}}{\mu_0}; \tag{2}$$
$$\vec{D} \equiv \vec{B}_e - \vec{M}_{efr}; \quad \vec{H} \equiv \vec{H}_e + \frac{\vec{M}_{mfr}}{\mu_0}.$$

Where: $\mathbf{M}_{efr}$ and $\mathbf{M}_{mfr}$ are the electric and magnetic matter fields produced by the free electric charges; $\mathbf{B}_e$, which is called the *electric flux density*, is a vector field composed of **D** and $\mathbf{M}_{efr}$.; and $\mathbf{H}_e$, which is called the *magnetic field intensity*, is a part of the vector field **H** that is composed of $\mathbf{H}_e$ and $\mathbf{M}_{mfr}$.

Then, (1) can be written in a more symmetrical form [7] as:

$$\begin{aligned} &\nabla\cdot\vec{B}_m=0; && \nabla\cdot\vec{B}_e=0;\\ &\frac{\partial\vec{B}_m}{\partial t}=-\nabla\times\vec{E}; && \frac{\partial\vec{B}_e}{\partial t}=\nabla\times\vec{H}_e. \end{aligned} \tag{3}$$

Where:

$$\begin{aligned} &\vec{B}_e=\vec{D}+\vec{M}_{efr}\equiv\left(\varepsilon_0\vec{E}+\vec{M}_{eb}\right)+\vec{M}_{efr}=\varepsilon_0\vec{E}+\left(\vec{M}_{eb}+\vec{M}_{efr}\right);\\ &\mu_0\vec{H}_e=\mu_0\vec{H}-\vec{M}_{mfr}\equiv\left(\vec{B}_m-\vec{M}_{mb}\right)-\vec{M}_{mfr}=\vec{B}_m-\left(\vec{M}_{mb}+\vec{M}_{mfr}\right). \end{aligned} \tag{4}$$

$\mathbf{M}_{eb}$ and $\mathbf{M}_{mb}$ are the electric and magnetic matter fields produced by the bound charges.

It is important to note that our definition of *electric flux density* $\mathbf{B}_e$ **is different** from the definition of IEV 121-11-40, where they call "electric flux density" to the *displacement* **D (**saying that this last terminology is obsolete**)**, despite not being divergence-less. Also, our definition of *magnetic field intensity* $\mathbf{H}_e$ **is different** from the definition of **H,** the *magnetic field strength* of IEV 121-11-56.

Equations (3) constitute the formulation of the electromagnetic theory that is called "Symmetrical theory of electromagnetism" [7].
The two vector equations in (3) can be interpreted as state equations, with the flux fields $\mathbf{B}_e$ and $\mathbf{B}_m$ representing the *electric state* and the *magnetic state*, respectively; and the intensity fields (**E** and $\mathbf{H}_e$**)** being the necessary inputs to produce a change in the states.

Using (4), (1) and (3) can be written as:

$$\begin{aligned} &\nabla\cdot\vec{B}_m=0; && \nabla\cdot\varepsilon_0\vec{E}=\rho_{et};\\ &\nabla\times\vec{E}=-\frac{\partial\vec{B}_m}{\partial t}; && \nabla\times\frac{\vec{B}_m}{\mu_0}=\vec{J}_{et}+\varepsilon_0\frac{\partial\vec{E}}{\partial t}. \end{aligned} \tag{5}$$

Where:

$$\begin{aligned} &\vec{M}_e\equiv\vec{M}_{efr}+\vec{M}_{eb};\quad \vec{M}_m\equiv\vec{M}_{mfr}+\vec{M}_{mb};\\ &\rho_{et}\equiv-\nabla\cdot\vec{M}_e;\\ &\vec{J}_{et}\equiv\frac{\partial\vec{M}_e}{\partial t}+\nabla\times\frac{\vec{M}_m}{\mu_0}. \end{aligned} \tag{6}$$

From (5) and (6) it is clearly seen that the sources of **E** and $\mathbf{B}_m$ are $\mathbf{M}_e$ and $\mathbf{M}_n$. Equations (5) constitute the Feynman's formulation of the electromagnetic theory [8].

If instead of separating the matter fields $\mathbf{M}_e$ and $\mathbf{M}_n$ into fields produced by free and bound charges, as in (6), we make the separation of the matter fields into fields produced by charges that are internal ($\mathbf{M}_{ei}$ and $\mathbf{M}_{mi}$) or external ($\mathbf{M}_{eex}$ and $\mathbf{M}_{mex}$) to the piece of matter considered; then we have:

$$\begin{aligned} &\nabla\cdot\vec{B}_m=0; && \nabla\cdot\vec{D}'=\rho_{ex};\\ &\nabla\times\vec{E}=-\frac{\partial\vec{B}_m}{\partial t}; && \nabla\times\vec{H}'=\vec{j}_{ex}+\frac{\partial\vec{D}'}{\partial t}. \end{aligned} \tag{7}$$

Where:

$$\begin{aligned} &\vec{M}_e\equiv\vec{M}_{eex}+\vec{M}_{ei};\quad \vec{M}_m\equiv\vec{M}_{mex}+\vec{M}_{mi};\\ &\rho_{ex}\equiv-\nabla\cdot\vec{M}_{eex};\\ &\vec{j}_{ex}\equiv\frac{\partial\vec{M}_{eex}}{\partial t}+\nabla\times\frac{\vec{M}_{mex}}{\mu_0};\\ &\vec{D}'\equiv\varepsilon_0\vec{E}+\vec{M}_{ei};\quad \vec{H}'\equiv\frac{\vec{B}_m-\vec{M}_{mi}}{\mu_0}. \end{aligned} \tag{8}$$

Equations (7) constitute the Landau & Lifshitz's formulation of the electromagnetic theory [9].

All the usual formulations of the electromagnetic theory, as (1), (3), (5) or (7), have in common the equations of the left-hand side involving the vector fields **E** and $\mathbf{B}_m$.
Usually, these left-hand side equations are solved in terms of the so-called *magnetic vector potential* "$\mathbf{A}_m$" and *electric scalar potential* "Φ":

$$\begin{aligned} &\nabla\cdot\vec{B}_m=0\Rightarrow\vec{B}_m=\nabla\times\vec{A}_m;\\ &\frac{\partial\vec{B}_m}{\partial t}=-\nabla\times\vec{E}\Rightarrow\nabla\times\left(\vec{E}+\frac{\partial\vec{A}_m}{\partial t}\right)=\vec{0}\rightarrow\vec{E}=-\left(\nabla\phi_m+\frac{\partial\vec{A}_m}{\partial t}\right). \end{aligned} \tag{9}$$

In most electromagnetic theories, the potentials have not physical reality, then the divergent of the magnetic potential can be arbitrarily chosen. The most common choices are:

$$\begin{aligned} &\nabla\cdot\left(\frac{\vec{A}_m}{\mu_0}\right)+\frac{\partial}{\partial t}\left(\varepsilon_0\phi_m\right)=0 \quad Lorenz'\,gauge;\\ &\nabla\cdot\vec{A}_m=0 \qquad\qquad\qquad Coulomb'\,gauge;\\ &\nabla\cdot\left(\frac{\vec{A}_m}{\mu}\right)+\frac{\partial}{\partial t}\left(\varepsilon\phi_m\right)=0 \quad Maxwell'\,gauge. \end{aligned} \tag{10}$$

In the "Symmetrical theory of electromagnetism" [7], there exist potentials that have physical reality, which are those related to the Hertz' potentials, which fulfill the following restriction:

$$\nabla\cdot\left(\frac{\vec{A}_m}{\mu_0}\right)+\frac{\partial}{\partial t}\left(\varepsilon_0\phi_m\right)=g_e c\approx\approx 0. \tag{11}$$

Saying that the Lorenz' gauge has physical reality. However, as in the normal Lorenz' gauge, there exist a manifold of magnetic potentials that are mathematically equivalent to the real potentials, in the sense that they produce the same magnetic flux density and the same electric field strength, which are the quantities that are universally recognized as having physical reality. These are:

$$\begin{aligned} &\vec{A}_m{}^*=\vec{A}_m+\nabla\chi;\quad \phi_m{}^*=\phi_m-\frac{\partial\chi}{\partial t};\\ &\vec{B}_m=\nabla\times\vec{A}_m=\nabla\times\vec{A}_m{}^*;\\ &\vec{E}=-\nabla\left(\phi_m-\frac{\partial\chi}{\partial t}\right)-\frac{\partial}{\partial t}\left(\vec{A}_m+\nabla\chi\right)=-\nabla\phi_m-\frac{\partial\vec{A}_m}{\partial t};\\ &With\quad \nabla^2\chi-\mu_0\varepsilon_0\frac{\partial^2\chi}{\partial t^2}=0. \end{aligned} \tag{12}$$

In the macroscopic formulation of the electromagnetic theory, usually the vector fields appearing in the right-hand side of (1), (3), (5) or (7) are expressed in terms of the fields **E** and $\mathbf{B}_m$, by means of the so-called *constitutive laws*.

For example, for the Maxwell-Hertz's formulation (see (1)), we have:

$$\vec{J}_{efr} = \sigma \vec{E}; \quad \vec{D} = \varepsilon \vec{E}; \quad \vec{H} = \frac{\vec{B}_m}{\mu}. \tag{13}$$

The parameters "σ", "ε" and "μ", appearing in (13), are called *electric conductivity*, *electric permittivity* and *magnetic permeability*, respectively.

Then, (1) can be written in terms of the potentials as:

$$\begin{aligned} &\vec{B}_m = \nabla \times \vec{A}_m; \quad \vec{E} = -\left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right); \\ &\nabla \cdot \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) = -\rho_{efr}; \\ &\nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu} \right) = -\sigma \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right). \end{aligned} \tag{14}$$

From the last two equations of (14) we can obtain:

$$\begin{aligned} &\nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu} \right) = -\sigma \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) \Rightarrow \\ &\frac{\partial}{\partial t} \left(\nabla \cdot \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) \right) = -\nabla \cdot \left(\sigma \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) = -\frac{\partial \rho_{efr}}{\partial t}. \end{aligned} \tag{15}$$

The last two equations of (14) are four scalar equations to determine the components of the four-vector formed by "$\mathbf{A}_m$" and "Φ/c" [10], which we will call "A_{em}", whose source is the free electric charge density. Equations (15) says that for a conduction dominated medium the real source is the time variation of the free electric charge density.
Then, in the case where the matter is described via *constitutive laws*, the field "A_{em}", with only four degrees of freedom, fully describes the electromagnetic field.
This is the formulation applied in [22].

When dealing with interference problems is very important to distinguish what belongs to the system being studied and what is considered an externally applied electromagnetic field.

Then, we will write the equations corresponding to (13) and (14) for the Landau & Lifshitz's formulation (see (7)), which makes this separation. In this case we have:

$$\frac{\partial \vec{D}'}{\partial t} = \sigma^* \vec{E} + \frac{\partial}{\partial t}\left(\varepsilon \vec{E} \right); \quad \vec{H}' = \frac{\vec{B}_m}{\mu^*}. \tag{16}$$

The parameters "σ*", "ε" and "μ*", appearing in (16), are also called *electric conductivity*, *electric permittivity* and *magnetic permeability*, respectively.

Then, (7) can be written as:

$$\begin{aligned} &\vec{B}_m = \nabla \times \vec{A}_m; \quad \vec{E} = -\left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right); \\ &\nabla \times \left(\frac{\nabla \times \vec{A}_m}{\mu^*} \right) = \vec{j}_{ex} - \sigma^* \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) - \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right); \\ &\nabla \cdot \left(\sigma^* \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) + \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) \right) = -\frac{\partial \rho_{ex}}{\partial t} = \nabla \cdot \vec{j}_{ex}. \end{aligned} \tag{17}$$

Again, the last two equations of (17) are four scalar equations to determine the components of the four-vector "A_{em}", formed by "$\mathbf{A}_m$" and "Φ/c", whose sources are the external electric charge density and the external current density. Then, also in this case the field "A_{em}", with only four degrees of freedom, fully describes the electromagnetic field.

We can write the last two equations of (17) in a different manner:

$$\begin{aligned} &\mu^* \varepsilon \frac{\partial^2 \vec{A}_m}{\partial t^2} + \mu^* \left(\sigma^* + \frac{\partial \varepsilon}{\partial t} \right) \frac{\partial \vec{A}_m}{\partial t} - \nabla^2 \vec{A}_m - \left(\frac{\nabla \mu^*}{\mu^*} \right) \times \left(\nabla \times \vec{A}_m \right) = \\ &= \mu^* \vec{j}_{ex} - \mu^* \left(\sigma^* + \frac{\partial \varepsilon}{\partial t} \right) \nabla\phi + \nabla\left(\mu^* \varepsilon \right) \frac{\partial \phi}{\partial t} - \nabla \left(\nabla \cdot \vec{A}_m + \mu^* \varepsilon \frac{\partial \phi}{\partial t} \right); \\ &\nabla \cdot \left(\sigma^* \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) + \frac{\partial}{\partial t} \left(\varepsilon \left(\nabla\phi + \frac{\partial \vec{A}_m}{\partial t} \right) \right) \right) = -\frac{\partial \rho_{ex}}{\partial t}. \end{aligned} \tag{18}$$

Equations (18) show that, in the case of the electromagnetic field within the matter, even in the Maxwell's gauge, in general, the equations for ϕ and $\mathbf{A}_m$ are not separable, as in the case of the vacuum. Besides that, the simple retarded solutions are not applicable.

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (17) is simplified, by the fact that we can express the external charge density in terms of the divergence of the current density, and for the last two equations of (17) we have:

$$\begin{aligned} &\nabla \cdot \left(\varepsilon \left(\nabla \Phi_\omega + i\omega \vec{A}_{m\omega} \right) \right) = -i \frac{\nabla \cdot \vec{j}_{ex\omega}}{\omega} - \nabla \cdot \left(\sigma^* \left(\vec{A}_{m\omega} - i \frac{\nabla \Phi_\omega}{\omega} \right) \right); \\ &\nabla \times \left(\frac{\nabla \times \vec{A}_{m\omega}}{\mu^*} \right) = \left(\vec{j}_{ex\omega} - \sigma^* \nabla \Phi_\omega \right) - i\omega \left(\sigma^* \vec{A}_{m\omega} + \varepsilon \nabla \Phi_\omega \right) + \omega^2 \varepsilon \vec{A}_{m\omega}. \end{aligned} \tag{19}$$

Equations (19) show clearly that the sources of the induced electromagnetic field are the external electric charge density and the external current density.
This is the formulation applied in [23,24].

3. Transmission-Lines

As in reference [1], and our (20) to (24) are the same than (11) to (15) of reference [1], in the case of the transmission lines, matter is predominantly distributed along a certain direction in space, which we will call it "z", making this a preferential direction.
As the four-dimensional vector field "A_{em}" transforms as a four-vector [10], for a transformation of coordinates between a system moving with a velocity "v", along the preferential z-axis, relative to another "rest" system, which we denote by the index "0", we have:

$$\begin{bmatrix} \frac{\phi}{c} \\ A_{mx} \\ A_{my} \\ A_{mz} \end{bmatrix} = \begin{bmatrix} \gamma\left(\frac{\phi_0}{c} - \frac{v}{c} A_{m0z} \right) \\ A_{m0x} \\ A_{m0y} \\ \gamma\left(A_{m0z} - \frac{v}{c} \frac{\phi_0}{c} \right) \end{bmatrix}; \quad \gamma \equiv \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}. \tag{20}$$

From (20) we can see that, for horizontal (along the z-axis) ideal transmission lines (having an uniform cross section), we can divide the electromagnetic field into two independent modes, each one with two degrees of freedom: a first one composed of "Φ" and "A_{mz}", which will be called *longitudinal mode*; and a second one composed of "A_{mx}" and "A_{my}", which will be called *transversal mode*.

Transmission lines constituted of insulated filamentary conductors, disposed along the z-axis, can internally produce only the longitudinal mode. The transversal mode must be externally produced.
The transversal mode can only be internally produced if the transversal extension of the conductive matter is relevant or if the filamentary conductors are not insulated.

For the longitudinal mode ($A_{mx} = A_{my} = 0$), from the first two equations of (11) or (14), we have:

$$E_x = -\frac{\partial \Phi}{\partial x};\quad E_y = -\frac{\partial \Phi}{\partial y};\quad E_z = -\frac{\partial \Phi}{\partial z} - \frac{\partial A_{mz}}{\partial t};$$
$$B_{mx} = \frac{\partial A_{mz}}{\partial y};\quad B_{my} = -\frac{\partial A_{mz}}{\partial x};\quad B_{mz} = 0. \tag{21}$$

Equations (21) say that the longitudinal mode is a TM mode, and, that the transversal electric field is *conservative* and is equal to the transversal gradient of the scalar potential Φ. This allows the definition of *transverse voltages* that are single valued [11, 12, 14].

For the transversal mode ($\Phi = A_{mz} = 0$), we have:

$$E_x = -\frac{\partial A_{mx}}{\partial t};\quad E_y = -\frac{\partial A_{my}}{\partial t};\quad E_z = 0;$$
$$B_{mx} = -\frac{\partial A_{my}}{\partial z};\quad B_{my} = \frac{\partial A_{mx}}{\partial z};\quad B_{mz} = \left(\frac{\partial A_{my}}{\partial x} - \frac{\partial A_{mx}}{\partial y}\right). \tag{22}$$

Equations (22) say that the transversal mode is a TE mode, with a non-conservative transversal electric field.

As an example of a pure longitudinal mode, we will see first the special case treated by Schelkunoff [11], of an infinite horizontal hollow conductor of arbitrary cross-section, rigid and made of perfectly conducting matter. Here, if we can assume that "Φ" and "A_{mz}" are separable, we have:

$$\Phi = -T(x,y)V(z,t);\quad \frac{A_{mz}}{\mu_0} = -T(x,y)I(z,t). \tag{23}$$

Where, in equations (23) the T functions are dimensionless functions of x and y.
Using (23) in (21), we have the Schelkunoff's TM modes [11]:

$$E_x = \frac{\partial T}{\partial x}V;\quad E_y = \frac{\partial T}{\partial y}V;\quad E_z = T\left(\frac{\partial V}{\partial z} + \mu_0 \frac{\partial I}{\partial t}\right);$$
$$B_{mx} = -\mu_0 \frac{\partial T}{\partial y}I;\quad B_{my} = \mu_0 \frac{\partial T}{\partial x}I;\quad B_{mz} = 0. \tag{24}$$

Also, from the equations on the right-hand side of (1) and (3), we have:

$$\left(\nabla \times \vec{H}\right)_x = -\frac{\partial}{\partial z}\left(\frac{\partial T}{\partial x}I\right) = -\frac{\partial T}{\partial x}\frac{\partial I}{\partial z} = \varepsilon \frac{\partial}{\partial t}\left(\frac{\partial T}{\partial x}V\right) = \varepsilon \frac{\partial T}{\partial x}\frac{\partial V}{\partial t};$$
$$\left(\nabla \times \vec{H}\right)_y = \frac{\partial}{\partial z}\left(-\frac{\partial T}{\partial y}I\right) = -\frac{\partial T}{\partial y}\frac{\partial I}{\partial z} = \varepsilon \frac{\partial}{\partial t}\left(\frac{\partial T}{\partial y}V\right) = \varepsilon \frac{\partial T}{\partial y}\frac{\partial V}{\partial t}; \tag{25}$$
$$\rightarrow \frac{\partial I}{\partial z} + \varepsilon \frac{\partial V}{\partial t} = 0.$$

Equations (24) and (25) display the main characteristic of Schelkunoff's TM modes, which is the fact that, even in the case of a hollow conductor made of perfectly conducting matter, despite having an electric field strength value equal to zero within the conductor, there can be an electromagnetic field inside the conductor, which propagates along the z axis, following transmission-line equations. The existence of this longitudinal mode and its predominance, even for hollow conductors that are slowly bent, leads to the theory of waveguides.

The other very important special case, where the longitudinal mode is predominant, is the case of a multi-conductor transmission line having, in general, many (N+1) filamentary conductors.
Here, following the same line of reasoning and adopting the terminology utilized in reference [1], we will describe the interaction of an arbitrary external electromagnetic field with a straight segment of a multiconductor line.
In Annex 1 the case of an elementary single-wire line is treated in detail. There we present the derivation of a general theory of the coupling of an external electromagnetic field to a conductor line and a detailed comparison of this generalized theory with the most important classical theories. This theory can be applied to different problems, such as, electromagnetic neural stimulation [16-20], the calculation of lightning-induced voltages [13].

In the case of a multiconductor line, we can obtain from (9) or (21) the value of the longitudinal electric field strength, **at a point internal to the "j" conductor** (see Figure 1), for j = 0 … N:

$$-\frac{\partial \Phi^j}{\partial z} - \frac{\partial A^j_{mz}}{\partial t} = E_z\left(x_j, y_j, z, t\right). \tag{26}$$

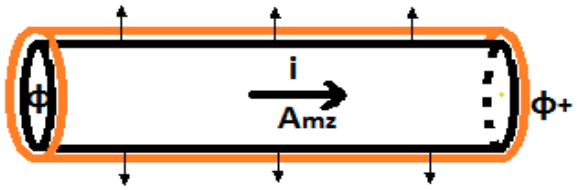


Figure 1 – Segment of a conductor, of length "Δz", between two scalar potential nodes.

Also, integrating the divergence of the last equation of (14) in a closed surface that involves **a segment of length "Δz" of the horizontal "j" conductor (**see Figure 2), we have:

$$\frac{\partial i^j}{\partial z} + \frac{\partial \lambda^j}{\partial t} = -I^j_{LS}. \tag{27}$$

Where, "i^j" is the total current flowing through the cross-section of the "j" conductor, "λ^j" is the charge accumulated on the surface of the "j" conductor, per unit length, and "$I_{LS}{}^j$" is the conduction current flowing out of the "j" conductor through the lateral surface, per unit length.

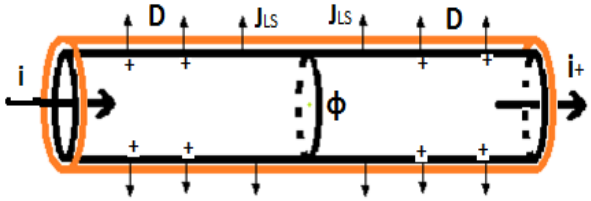


Figure 2 – Segment of a conductor, of length "Δz", around a scalar potential node.

If, as usual, we assume that for this kind of transmission line, for which the distances among the different conductors of the line is much shorter than its length, the Maxwell's concept of inductance and capacitance coefficients are valid [25,26].

This will allow us to express the magnetic potential in a short segment of a filamentary conductor, at a point "z" along the line, in terms of the current in all the conductors at the same point "z", and, the charge at the surface of a conductor, at a point "z", in terms of the scalar potential at all the conductors at the same point "z". Then, we have:

$$\begin{aligned} &A_{mz}^{j} = A_{mz}^{j(\Delta)} + A_{mz}^{j(ext)}; \qquad \Phi^{j} = \Phi^{j(\Delta)} + \Phi^{j(ext)}; \\ &A_{mz}^{j(\Delta)} = \left[L_k^{j^*} \right] i^k \left(z,t \right); \quad \lambda^j = \left[C_k^{j^*} \right] \Phi^{k(\Delta)} \left(z,t \right). \end{aligned} \tag{28}$$

Where: "$A_{mz}^{j(\Delta)}$" and "$\Phi^{j(\Delta)}$" are the potentials produced by the matter existing in all the conductors within the segment "Δz", at the point "z"; and "$A_{mz}^{j(ext)}$" and "$\Phi^{j(ext)}$", are the potentials produced by the matter existing in all the conductors, outside the segment "Δz", plus the potentials "$A_{mz}^{j(ex)}$" and "$\Phi^{j(ex)}$" representing the externally applied field.
Using Ohm's law for the line conductors, we have:

$$E_z \left(x_j, y_j, z, t \right) = R_{(cj)} i^j \left(z,t \right). \tag{29}$$

Where "$R_{(cj)}$" is the resistance, per unit length, of the conductor "j".

Extracting from the conduction current "I_{LS}^{j}", flowing out of the conductor through the lateral surface, the part that is due to the linear leakage current, we can write:

$$I_{LS}^{j} \left(z,t \right) = \left[G_k^{j^*} \right] \Phi^{k(\Delta)} + I_{LS}^{j^*} \left(z,t \right). \tag{30}$$

Then, (26) and (27) can be written as:

$$\frac{\partial \Phi^j}{\partial z} + \left(R_{(cj)} i^j + \left[L_k^{j^*} \right] \frac{\partial i^k}{\partial t} \right) = - \frac{\partial A_{mz}^{j(ext)}}{\partial t}; \tag{31}$$

and

$$\frac{\partial i^j}{\partial z} + \left[G_k^{j^*} \right] \Phi^{k(\Delta)} + \left[C_k^{j^*} \right] \frac{\partial \Phi^{k(\Delta)}}{\partial t} = - I_{LS}^{j^*}. \tag{32}$$

Where: "[L*]", is the matrix of the inductance coefficients, per unit length; "[C*]" is the matrix of the capacitance coefficients, per unit length; and "[G*]" is the matrix of conductance coefficients, per unit length, between the segments of all conductors at the point "z".

To describe a rectilinear transmission line of a uniform cross-section and of a finite length, we can use here again, as usual in transmission-line theory [12], the matrices [L], [C] and [G], which are calculated neglecting the retardation effects and when every conductor of the line is uniformly charged with a charge density, per unit length, which is equal to its value at the point "z", and the current in all segments along the line, in any conductor, is equal to its value at the point "z". The potentials produced in this situation are the *static potentials* "$A_{mz}^{(st)}$" and "$\Phi^{(st)}$". Using for every conductor "j" the following definitions, we have:

$$\begin{aligned} &A_{mz}^{j(\text{int})} \equiv A_{mz}^{j} - A_{mz}^{j(ex)}; \quad A_{mz}^{j(c)} \equiv A_{mz}^{j(\text{int})} - A_{mz}^{j(st)}; \\ &\rightarrow A_{mz}^{j} = A_{mz}^{j(st)} + \left(A_{mz}^{j(ex)} + A_{mz}^{j(c)} \right); \\ &A_{mz}^{j(st)} \equiv \left[L_k^{j} \right] i^k \left(z,t \right); \\ &\Phi^{j(\text{int})} \equiv \Phi^{j} - \Phi^{j(ex)}; \qquad \Phi^{j(c)} \equiv \Phi^{j(\text{int})} - \Phi^{j(st)}; \\ &\rightarrow \Phi^{j} = \Phi^{j(st)} + \left(\Phi^{j(ex)} + \Phi^{j(c)} \right); \\ &\lambda^j \left(z,t \right) \equiv \left[C_k^{j} \right] \Phi^{k(st)}; \quad \left[G_k^{j^*} \right] \Phi^{k(\Delta)} \equiv \left[G_k^{j} \right] \Phi^{k(st)}. \end{aligned} \tag{33}$$

Where the usual summation rule over repeated index is applied.

Then, we can write (31) and (32) as:

$$\frac{\partial \Phi^j}{\partial z} + \left(R_{(cj)} i^j + \left[L_k^{j} \right] \frac{\partial i^k}{\partial t} \right) = - \frac{\partial \left(A_{mz}^{j(ex)} + A_{mz}^{j(c)} \right)}{\partial t}; \tag{34}$$

and

$$\begin{aligned} &\frac{\partial i^j}{\partial z} + \left[G_k^{j} \right] \Phi^k + \left[C_k^{j} \right] \frac{\partial \Phi^k}{\partial t} = \\ &= \left[G_k^{j} \right] \left(\Phi^{k(ex)} + \Phi^{k(c)} \right) + \left[C_k^{j} \right] \frac{\partial \left(\Phi^{k(ex)} + \Phi^{k(c)} \right)}{\partial t} - I_{LS}^{j^*}. \end{aligned} \tag{35}$$

As mentioned in reference [1], "the main variables in (34) and (35) are the scalar and the magnetic potentials, which are not uniquely determined (see (12)). Their values, as well as the value of the inductance and capacitance coefficients, are affected by the choice of the reference point, which is the point where the value of the potentials is equal to zero.
But, as the fields **E** and $\mathbf{B}_m$ are uniquely determined and, for practical purposes, the important voltages are the ones occurring between the conductors of the line, this fact is of little or no importance."

We will choose the reference for the potentials at the infinity, or at a line located very far from the multiconductor transmission line.
With this choice of reference, (34) can be interpreted as the application of Kirchhoff's circuits law applied to a "mesh" formed by a segment of the conductor "j" of the line and the reference line; and (35) as the application of Kirchhoff's circuits law applied to a potential node on the conductor "j" of the line. Then, we can see that the longitudinal TM mode is the one described by circuit theory.

Also, (34) and (35) are completely general and rigorous equations that describe the interaction of an external electromagnetic field with the considered filamentary multi-conductor transmission line, being the only assumptions in deriving these equations: the thin-wire approximation for all the conductors and the validity of Ohm's law for all the conductors. Thus, they constitute a generalized formulation of the electromagnetic coupling to a transmission line that, under the proper simplifications, reduces to the standard coupling theories [12-14,21] (see Annex 2).

One of the advantages of the time domain formulation of the longitudinal mode is that (34) and (35) can also be written as state equations:

$$\frac{\partial i^j}{\partial t} = - \left[L_j^k \right]^{-1} \left(\frac{\partial \Phi^k}{\partial z} + R_{(ck)} i^k \right) - \left[L_j^k \right]^{-1} \left(\frac{\partial \left(A_{mz}^{k(ex)} + A_{mz}^{k(c)} \right)}{\partial t} \right); \tag{34a}$$

and

$$\begin{aligned} &\frac{\partial \Phi^j}{\partial t} = - \left[C_j^k \right]^{-1} \left(\frac{\partial i^k}{\partial z} + \left[G_m^k \right] \Phi^m \right) + \\ &+ \left(\frac{\partial \left(\Phi^{j(ex)} + \Phi^{j(c)} \right)}{\partial t} + \left[C_j^k \right]^{-1} \left[G_m^k \right] \left(\Phi^{m(ex)} + \Phi^{m(c)} \right) - \left[C_j^k \right]^{-1} I_{LS}^{k*} \right). \end{aligned} \tag{35a}$$

Then, (34a) and (35a), which are a generalization of (21a) and (22a) of reference [1], as already mentioned in reference [1], "can be interpreted as saying that the scalar potential "Φ^j" and the current "i^j" represent the time evolving state of the system, at a point internal to the "j" conductor." **Showing that the essence of the circuit theory is to assume that the longitudinal mode is predominant.**

Equations (34a) and (35a), after being spatially discretized into N segments (for n = 1...N), can be written as:

$$\frac{di_n^j}{dt} = -\left[L_j^k\right]^{-1}\left(R_{(ck)}i_n^k\right) - \left[L_j^k\right]^{-1}\frac{\left(\Phi_{n+1}^k - \Phi_n^k\right)}{\Delta z} + \\ -\left[L_j^k\right]^{-1}\left(\frac{\partial\left(A_{mzn}^{k(ex)} + A_{mzn}^{k(c)}\right)}{\partial t}\right); \tag{34b}$$

and

$$\frac{d\Phi_n^j}{dt} = -\left[C_j^k\right]^{-1}\left[G_m^k\right]\Phi_n^m - \left[C_j^k\right]^{-1}\frac{\left(i_n^k - i_{n-1}^k\right)}{\Delta z} + \\ + \begin{pmatrix} \frac{\partial\left(\Phi_n^{j(ex)} + \Phi_n^{j(c)}\right)}{\partial t} + \left[C_j^k\right]^{-1}\left[G_m^k\right]\left(\Phi_n^{m(ex)} + \Phi_n^{m(c)}\right) + \\ -\left[C_j^k\right]^{-1} I_{LSn}^{k*} \end{pmatrix}. \tag{35b}$$

Equations (34b) and (35b), as also mentioned in reference [1], "are a system of coupled ordinary differential equations (ODE), for the scalar potential at potential nodes and the current at current nodes." Also, as already mentioned in reference [15], "this system of coupled ordinary differential equations can then be solved using the powerful ODE solvers now existing" [27].

The main advantage of describing uniform lines by per-unit length parameters, calculated using the static potentials is that these parameters are constant and easy to calculate, particularly, when the reference is taken at the infinity, because it is a reference independent of the direction of the line.
For obtaining approximate solutions of non-uniform lines, assuming that the longitudinal mode is predominant, the non-uniform line is usually modeled as a cascade of uniform sections, conductively connected [28]; or in the case of lines with periodically varying cross-section, such as the cables composed of twisted-wire pairs, the line is modeled as an equivalent line, having per-unit length parameters equal to the average over the period [29].

Up to now, we have studied the uniform or slowly non-uniform part of the line.
As mentioned in reference [1], using reference [15], "we will take advantage of this formulation that allows to include the interaction of a segment of conductor with any known arbitrary external electromagnetic field, which can be described by $\Phi^{(ex)}$ and $\mathbf{A}^{(ex)}$, to include the terminations of the line, the discontinuities or even conductors attached perpendicular to the line, such as, groundings or other equipment connected.".
Considering that the longitudinal mode is predominant, we can model them as circuit elements located between two potential nodes, which are displaced in space along a certain direction.

For a vertical conductor (y axis direction) that is relatively small (compared to the minimum wavelength of interest), so that we can neglect the corrections due to the time delay in the production of the potentials, and neglecting also the non-linear transversal conduction current I_{LS}^*, we have in a manner analogous to (31) and (32):

$$\frac{\partial\left(\Phi - \Phi^{(ex)}\right)}{\partial y} + R_c i + L\frac{\partial i}{\partial t} = -\left(\frac{\partial\Phi^{(ex)}}{\partial y} + \frac{\partial A_{my}^{(ex)}}{\partial t}\right); \tag{36}$$

and

$$\frac{\partial i}{\partial y} + G\left(\Phi - \Phi^{(ex)}\right) + C\frac{\partial\left(\Phi - \Phi^{(ex)}\right)}{\partial t} = 0. \tag{37}$$

This is the explanation of the important remark quoted in the conclusions of reference [14] "When using the scattered-voltage formulation, it must be remembered that the vertical component of the incident electric field appears as a voltage source in the line terminations."

If, like in the case of a grounding of the reference conductor, the vertical conductor is connected at the potential node "Φ_n^0", the modified equations for this node are:

$$\frac{di_n^0}{dt} = -\left[L_0^k\right]^{-1}\left(R_{(ck)}i_n^k\right) - \left[L_0^k\right]^{-1}\frac{\left(\Phi_{n+1}^k - \Phi_n^k\right)}{\Delta z} + \\ -\left[L_0^k\right]^{-1}\left(\frac{\partial\left(A_{mzn}^{k(ex)} + A_{mzn}^{k(c)}\right)}{\partial t}\right); \tag{38}$$

and

$$\frac{d\Phi_n^0}{dt} = -\left[C_0^k\right]^{-1}\left[G_m^k\right]\Phi_n^m - \left[C_0^k\right]^{-1}\frac{\left(i_n^{k1} + i_n^k - i_{n-1}^k\right)}{\Delta z} + \\ + \begin{pmatrix} \frac{\partial\left(\Phi_n^{0(ex)} + \Phi_n^{0(c)}\right)}{\partial t} + \left[C_0^k\right]^{-1}\left[G_m^k\right]\left(\Phi_n^{m(ex)} + \Phi_n^{m(c)}\right) + \\ -\left[C_0^k\right]^{-1} I_{LSn}^{k*} \end{pmatrix}. \tag{39}$$

Where, in this case, only "i_n^{01}", which is the current in the first segment of the vertical conductor, is different from zero.

As mentioned in reference [1], "this formulation allows representing the groundings not as a connection to the ground, but as a connection to the grounding electrodes, whose potential is to be considered as one of the states of the system.
This is very convenient for the real cases where the ground is not perfectly conducting, and it is considered as the return conductor of the line.".

Finally, as in reference [1], to include the effect of the transversal mode, we will define the so-called *conductor voltages*.

Usually, in standard transmission-line theory, the *conductor voltages* are defined relative to one of the conductors chosen to be the reference [12, 30]. This was the choice indirectly adopted in reference [1].
Here, we will define the *conductor* "j" *voltage*, at a point "z" along the conductor, as the integral of the transversal electric field strength "$\mathbf{E}_t$" between the conductor "j" and the <u>reference of the potentials</u>:

$$V^j\left(z,t\right) \equiv -\int_{\infty}^{\left(x_j,y_j\right)}\vec{E}_t\cdot d\vec{l}_t = \\ = \Phi^j + \int_{\infty}^{\left(x_j,y_j\right)}\frac{\partial\vec{A}_{mt}^{(ex)}}{\partial t}\cdot d\vec{l}_t \equiv \Phi^j + V_{in}^j. \tag{40}$$

Where we will assume that "A_{mt}" is the externally produced field, but, as mentioned in reference [1], "the formulation also allows to include the transversal field produced by transversal conductors.". Then, we can modify (34) and (35), by means of (40), and obtain:

$$\frac{\partial V^j}{\partial z}+R_{(cj)}i^j+\left[L_k^j\right]\frac{\partial i^k}{\partial t}= \\ =-\frac{\partial}{\partial t}\left(A_{mz}^{j(ex)}-\int_{\infty}^{(x_j,y_j)}\frac{\partial \vec{A}_{mt}^{(ex)}}{\partial z}\cdot d\vec{l}_t\right)-\frac{\partial A_{mz}^{j(c)}}{\partial t}; \tag{41}$$

and

$$\frac{\partial i^j}{\partial z}+\left[G_k^j\right]V^k+\left[C_k^j\right]\frac{\partial V^k}{\partial t}=\left[G_k^j\right]V^{k(ex)}+\left[C_k^j\right]\frac{\partial V^{k(ex)}}{\partial t} \\ +\left(\left[G_k^j\right]\Phi^{k(c)}+\left[C_k^j\right]\frac{\partial \Phi^{k(c)}}{\partial t}-I_{LS}^{j*}\right). \tag{42}$$

Where, as mentioned in reference [1], "we can see that writing the equations in terms of the *conductor voltages* produce new terms in (41) and (42), due to the existence of induced voltages (integration path dependent) produced by the external transversal mode."

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (41) and (42) can be written as:

$$\frac{dV_\omega^j(z)}{dz}+\left(\left[R_k^j\right]+i\omega\left[L_k^j\right]\right)I_\omega^k(z)=S_{E\omega}^j(z); \tag{43}$$

and

$$\frac{dI_\omega^j(z)}{dz}+\left(\left[G_k^j\right]+i\omega\left[C_k^j\right]\right)V_\omega^k(z)=S_{H\omega}^j(z). \tag{44}$$

Which, without the sources, are the standard transmission-line equations in frequency domain, such as shown in (2) of reference [30] or (6.7) and (6.8) of reference [31].

As already mentioned in reference [1], "we must note that, for the longitudinal mode, the important values of the externally produced field are at points located inside the conductors, and its effect can be represented as a voltage or a current source; while, for the transversal mode, the important values of the externally produced field are at points located in the dielectric between the conductors, and its effect is to produce an induced voltage between the conductors.".

Also, as mentioned in reference [1], "it must be emphasized that, in order to define the concept of voltage associated with one point of the line, it is necessary to introduce a reference.
When the reference is at the infinity, or in the case of an imperfectly conducting Earth where the "Reference perfect conductor plane" is located at a remote position within the Earth, both the conductor "j" voltage and the scalar potential of the "j" conductor, will have embedded the complexities of the electromagnetic field distribution in the ground.
Calculate the electromagnetic field distribution within the Earth is a very difficult task, but fortunately, the voltages and the potential differences **between aerial conductors**, which are the ones of practical importance, will not depend directly on those complexities.".

4. Interferences produced by external Disturbances

In this part, as in part IV of reference [1], we will apply the generalized theory of the electromagnetic coupling to a transmission line, developed in part 3, to analyze the interference on a transmission line produced by external disturbances, which are commonly classified into conducted and radiated disturbances.

To study the interference due to an external disturbance, we must first separate the sources of the externally produced electromagnetic fields in two classes: the normal external sources and the disturbing external sources.

In the interconnections, which are multiconductor transmission lines, the dominant mode is the longitudinal mode, and, the longitudinal and the transversal modes are practically decoupled.
Then, as in reference [1], we will assume that the *normal operation mode*, which is driven by normal lumped external excitation sources, is a longitudinal mode.

IEV 161-03-27 says, for conducted disturbances, that the energy is transferred via one or more conductors. So, conducted disturbances are locally produced.
IEV 161-03-28 says, for radiated disturbances, that the energy is transferred through space in the form of electromagnetic waves; and, it notes that "The term "radiated disturbance" is sometimes used to cover induction phenomena". Then, radiated disturbances are always remotely produced.

Here, as in reference [1], "we will divide the externally produced disturbances in two classes: the first class that produce Φ and A_{mz}, which will be called *longitudinal mode disturbances*; and the second class that produce A_{mx} and A_{my}, which will be called *transversal mode disturbances*."

Also, as in reference [1], within the internally-produced longitudinal mode, "$\Phi^{(int)}$" and "$A_{mz}^{(int)}$", we must make a distinction between the scattered part and the part that is produced by the current that is injected by lumped external sources, both normal and disturbing.

Then, we have:

$$\Phi^j=\Phi^{j(\mathrm{int})}+\Phi^{j(ex)}\equiv\left(\Phi^{j(sc)}+\Phi^{j(exinjN)}+\Phi^{j(exinjD)}\right)+\Phi^{j(ex)}; \\ A_{mz}^j=A_{mz}^{j(\mathrm{int})}+A_{mz}^{j(ex)}\equiv\left(A_{mz}^{j(sc)}+A_{mz}^{k(exinjN)}+A_{mz}^{k(exinjD)}\right)+A_{mz}^{j(ex)}; \\ \Rightarrow i^j=\left[L_j^k\right]^{-1}\left(A_{mz}^{k(sc)}+A_{mz}^{k(exinjN)}+A_{mz}^{k(exinjD)}\right)=i^{j(sc)}+i^{j(exinj)}; \\ A_{mx}^j=A_{mx}^{j(ex)};\quad A_{my}^j=A_{my}^{j(ex)}. \tag{45}$$

Where: the index "N" indicates normal, the index "D" indicates disturbing and the index "exinj" indicates produced by the lumped external sources, conductively connected to the line, that injects current in the line. Our (45) is equal to (33) of reference [1].

In this paper, neglecting the internally-produced transversal magnetic potential, which is produced by the current in the transversal conductors and the transversal leakage currents, we have assumed, as in reference [1], that the transversal mode is produced only by the radiated disturbances.

As mentioned in reference [1], the effect of the transversal mode is:
- to produce an induced voltage between the conductors, and
- to act as a lumped voltage source located at the transversal conductors, such as the line terminations and groundings.

Then:
- The so-called "*conducted disturbances*" (IEV 161-03-27) are characterized by the fields $\Phi^{(exinjD)}$ and $A_{mz}^{(exinjD)}$, which are

produced by a current locally injected into the line, by lumped disturbing external sources conductively connected to it; and
- The so-called "*radiated disturbances*" (IEV 161-03-28) are characterized by the fields $\Phi^{(ex)}$, $A_{mz}^{(ex)}$, $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, which are produced by remotely located sources.

The confusing note existing in IEV 161-03-28 that "The term "*radiated disturbance*" is sometimes used to cover induction phenomena", can be explained because the electromagnetic field, produced by remotely located sources, includes both the induction field and the radiation field:

$$\begin{aligned} \Phi^{j(ex)} &= \Phi^{j(ind)}; \\ A_{mx}^{j(ex)} &= A_{mx}^{j(ind)} + A_{mx}^{j(rad)}; \\ A_{my}^{j(ex)} &= A_{my}^{j(ind)} + A_{my}^{j(rad)}; \\ A_{mz}^{j(ex)} &= A_{mz}^{j(ind)} + A_{mz}^{j(rad)}. \end{aligned} \tag{46}$$

Where: the index "*ind*" indicates *induction field*, which is negligible in the far field region, and the index "*rad*" indicates *radiation field*.

From equations (45) we can see that:
1- The conducted disturbances, "$\Phi^{(exinjD)}$" and "$A_{mz}^{(exinjD)}$", which produce scalar potentials and inject currents into the line, are longitudinal mode disturbances that affect only the longitudinal mode; while
2- The radiated disturbances, which are composed of longitudinal mode disturbances, $\Phi^{(ex)}$ and $A_{mz}^{(ex)}$, and transversal mode disturbances, $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, affect:
- The longitudinal mode, directly by the longitudinal mode disturbances through the scalar potential "$\Phi^{(ex)}$" and the magnetic potential "$A_{mz}^{(ex)}$" along the conductors; and indirectly by the transversal mode disturbances through the lumped voltage sources at the transversal conductors and terminations, which represent the effect of "$A_{mt}^{(ex)}$" along the transversal conductors;
- The transversal mode, through the magnetic potentials $A_{mx}^{(ex)}$ and $A_{my}^{(ex)}$, which produce an induced voltage between the conductors.

This is the reason why, only when the longitudinal is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.

Also, this explains why current injection and capacitive clamp testing methods represent only the effect of disturbances on the longitudinal mode.

Examples of radiated disturbances, where the longitudinal mode disturbances are dominant, can be seen in references [14,21,32], which deal with lightning-induced voltages; and in reference [33], which deal with voltages induced in twisted-wire pairs by a parallel wire excited by a voltage source.

Examples of radiated disturbances, where the transversal mode disturbances are important can be seen in:
- reference [14], where the exciting source is a lumped voltage source, which represents the induced voltage produced by the transversal mode disturbance, along a transversal conductor at input terminal of the line;
- reference [34], where the induced voltage, produced by the transverse mode disturbance, is an important part of the total voltage, and even the dominant part for the first three microseconds.

In reference [2], the conducted disturbances are defined as "Any deviation from the ideal voltage or current waveform", meaning the presence of a disturbing scalar potential or a longitudinal magnetic potential produced by the current, which is locally injected by lumped disturbing external sources.
Examples of common types and new types of power quality disturbances can be seen in reference [2].

5. Application Case Results

To show the usefulness of this formulation, we present, as in reference [1], the results of some application cases to real transmission lines with transversal conductors, which are connected to the earthing electrodes, in the case of a real ground that is not a perfect conductor. Adding here, in order to show the interpretation power of this formulation, some important practical and engineering conclusions that had gone unnoticed in other calculations made with previously proposed approaches/software tools.

Firstly, as in reference [1], we present in Figure 3 that corresponds to Figure 1 of reference [1], the results of the calculation of the phase-to-neutral voltage, induced by a vertical return stroke that strikes close to a line, which have a neutral wire grounded at only one point, with a grounding resistance of 100 Ω.

As mentioned in reference [1], the purpose of this example is to show that, in this case, the impinging electromagnetic pulse, composed of both a longitudinal mode disturbance and a transversal mode disturbance, initially practically produces only common mode on the line, and, the phase-to-neutral voltage is negligible until the instant when the electromagnetic disturbance reaches the grounding conductor of the neutral wire.

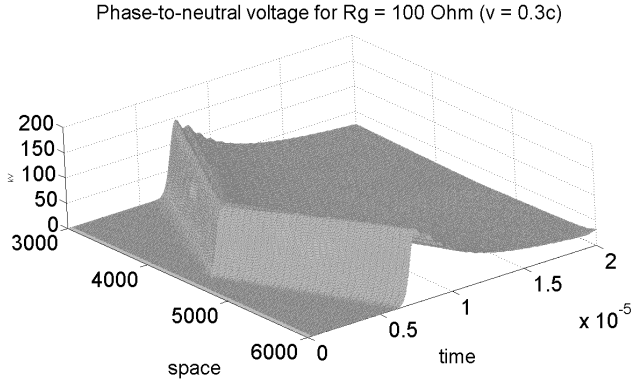


Figure 3 - Phase-to-neutral voltage induced by a 100 kA (2x40 μs) return stroke, with velocity 0.3c, which strikes at z = 4000 m, at 100 m from a line, with a neutral wire having an isolated grounding at z = 4500 m, with Rg = 100 Ω. Data taken from [15]

Then, due to the presence of the transversal grounding conductor of the neutral wire, when the electromagnetic transversal mode disturbance reaches the grounding conductor, the current produced by the impinging electric field along the conductor generates voltage reduction waves, which are different in size in the different conductors, thus producing a phase-to-neutral voltage pulse, which propagates in both directions along the line, representing the conversion of common mode into differential mode.

This effect that had gone unnoticed in other calculations made with previously proposed approaches/software tools [15,35], has the important engineering conclusion that the main mitigating effect for the induced voltages is not produced by the shielding

wire acting as an extended protective device but by the shielding wire grounding acting as a localized protective device.

As a second example, we present in Figure 4 that corresponds to Figure 2 of reference [1], the results of the calculation of the phase-to-neutral voltage, induced by a vertical return stroke that strikes close to a line, having a neutral wire that is periodically grounded.

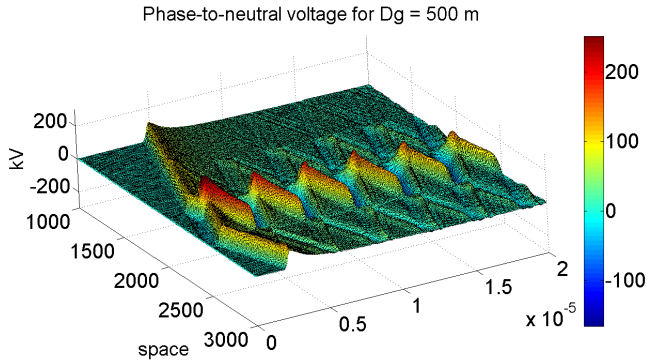


Figure 4 - Phase-to-neutral voltage, in a line periodically grounded each 500 m (Rg = 10 Ohms), produced by a 30kA return stroke (0.3x40 μs), occurring at z=2375 m, 50 m from the line. Data taken from [15].

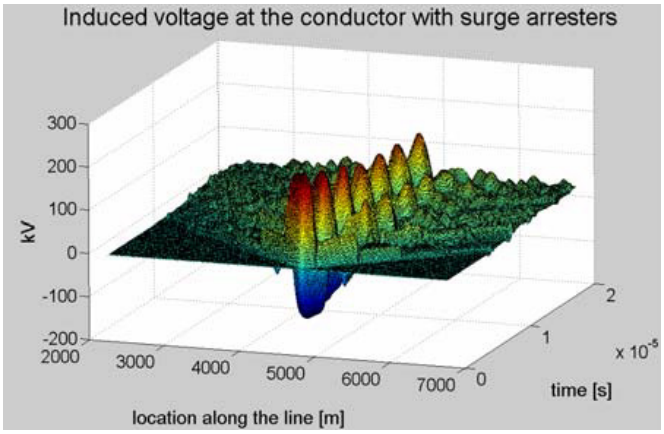
From Figure 4 it can be seen, as already observed in Figure 3, that the impinging electromagnetic pulse produces initially only a common mode perturbation on the line, which propagates in both directions starting from the point of the line which is closest to the stroke location, and the phase-to-neutral voltage is negligible until the instant when the electromagnetic disturbance reaches a transversal grounding conductor of the neutral wire. Then, a phase-to-neutral voltage pulse is produced, due to the injection of current in the transversal wire, which propagates in both directions along the line.

In this case, the effect of the presence of the transversal periodic grounding conductors is to confine the bigger phase-to-neutral overvoltage in the region of the line which is closest to the return stroke location.

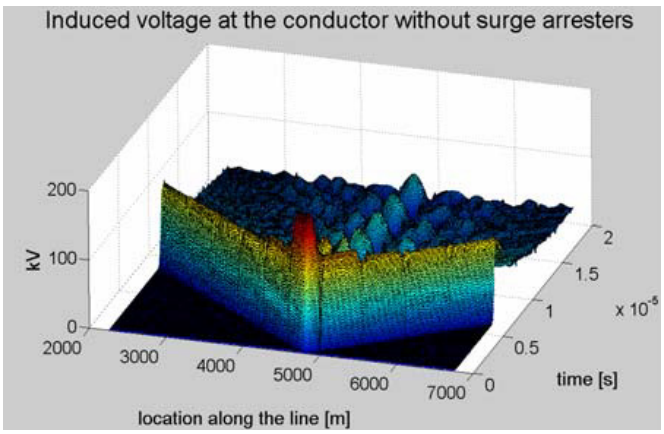
Also in this case, the calculations pointed to an effect, which had gone unnoticed in other calculations made with previously proposed approaches/software tools [36], that the presence of the grounding reduces the overvoltage at its location and behind it (seen from the return stroke location), and also it reduces the overvoltage in front of it, but only within a certain distance, fact which defines an "effective distance".
Then, as mentioned in reference [15], "the shielding wire groundings not only do not protect the line span in front of the return stroke location, but they also produce big positive phase-to-neutral overvoltage".

In the next example, we present in Figure 5 that corresponds to Figure 3 of reference [1], the results of the calculation of the voltage induced on the conductors of a line, by a vertical return stroke that strikes close to a line, having surge arresters placed periodically on one of the conductors.

From Figure 5, we can see that the effect of the presence of the transversal surge arresters, with its grounding conductors, is:
- On the conductor with surge arresters, to confine the phase-to-reference ground overvoltage, which is oscillating, in the region of the line close to the return stroke location, being the bigger overvoltage, in the region of the line closest to the return stroke location; and,
- On the conductor without surge arresters, to produce just a mild dampening of the overvoltage, produced by the impinging electromagnetic pulse, outside the region of the line closest to the return stroke location.



a) Voltage at the conductor with surge arresters.



b) Voltage at the conductor without surge arresters.

Figure 5– Voltages along a line with surge arresters on one of the conductors, at each 450 m (Rg = 10 Ohms), produced by a 45kA lightning discharge (2x40 μs) propagating with a velocity v = 0.3 c, which strikes at z =4800 m, at 70 m from the line. Data taken from [37].

As a result, they produce a conversion of common mode into differential mode overvoltage.
Also, in this case, the calculations pointed to an effect that had gone unnoticed in other calculations made with previously proposed approaches/software tools [36] and which has engineering importance. As mentioned in reference [37], "Multiple surge arresters, placed at periodic intervals along the line, protect the line outside the span in front of the lightning strike location; but the protection afforded to the span in front of the lightning strike location depends on the risetime of the induced voltages. Consequently, any statement on the effectiveness of an interval between arresters should be qualified by a risetime for which it is valid."

6. Summary and Conclusions

We have shown that, when the matter is macroscopically described by *constitutive laws*, the electromagnetic field within the matter can be fully described by the potentials: the *magnetic vector*

potential "$\mathbf{A}_m$" and the *electric scalar potential* "Φ", with its four degrees of freedom.
This conclusion is valid for any time scale and for any frequency, provided the *constitutive laws* are valid.

We have shown that, when the matter is predominantly distributed along a certain direction in space, as in a transmission line, the electromagnetic field can be divided into two practically independent modes, each one with two degrees of freedom: a *longitudinal mode* having "Φ" and "A_{mz}", and, a *transversal mode* having "A_{mx}" and "A_{my}".
Kirchhoff's laws and circuit theory applies to the two degrees of freedom of the longitudinal mode: the scalar potential "Φ" and the current "i" in the conductor, which is related to "A_{mz}" by the concept of inductance. They represent the time evolving state of the system, at points internal to the conductors.

Using the longitudinal mode as the fundamental building block, and if the longitudinal and the transversal modes are practically independent, we present the derivation of a generalized theory of the electromagnetic field coupling to a multiconductor line, in time domain, that, as usual, predicts the propagation of the scalar potential and the current along the line.
We have shown that this generalized coupling theory, under the proper simplifications, reduces to the standard coupling theories and the transmission-line equations there obtained also reduce to the standard transmission-line equations.

We have also shown that, when the longitudinal is predominant, the theory can be extended to include the terminations of the line, the discontinuities or even conductors attached perpendicular to the line.

Also, this formulation allows representing the groundings not as a connection to the reference ground, but as a connection to the grounding electrodes, whose potential is to be considered as one of the states of the system. This is very convenient for the real cases where the ground is not perfectly conducting.

Analyzing the interference produced by external disturbances in these terms, we have assumed that the normal operation mode, which is a differential mode driven by normal lumped external excitation sources, is a longitudinal mode.

We also have divided the externally produced disturbances in two classes: *longitudinal mode disturbances* and *transversal mode disturbances*.

We have shown that the *conducted disturbances* are longitudinal mode disturbances that affect only the longitudinal mode, and the *radiated disturbances* are composed of longitudinal mode disturbances and transversal mode disturbances, both of which affect the longitudinal mode.

We have shown that the transversal mode disturbances, which are part of the *radiated disturbances*, affect also the transversal mode. The transversal mode produces induced voltages (integration path dependent) in the insulation between the conductors of the line, and, between any conductor and the reference adopted.

This is the reason why, only when the longitudinal is predominant, and, the longitudinal and the transversal modes are practically decoupled, EMI can be simulated using circuit simulation software's.

In order to illustrate the usefulness of this interpretation, we have presented the results of the application of this formulation to real lines having perpendicular conductors attached to them, some of them connected to earthing electrodes. We have shown that the main effect of the presence of these conductors is to produce the conversion of common mode into differential mode.

Acknowledgment

The author would like to thank to LACTEC for supporting this work, and, to thank to the members of CIGRE WG C4.44, to his colleagues at LACTEC and to two anonymous reviewers for their comments and questions.

Annexes

1. Electromagnetic Field Coupling to a Conductor Line

Here, following the same line of reasoning utilized in part 3 and adopting the terminology utilized in reference [1], the very important special case of an infinite horizontal filamentary solid conductor submitted to an externally applied electromagnetic field, which has been studied since Maxwell's time, will be studied.

Historically, firstly it was studied the electromagnetic field produced by it, when the externally applied field impressed a sinusoidal current in the conductor [38]; but lately, it has been studied the scattered field produced by it, when excited by an externally impinging full-wave electromagnetic field (time-harmonic electromagnetic field) [13].

In this paper, we will study the case of a horizontal filamentary solid conductor, of an infinite extent, submitted to an arbitrary externally applied electromagnetic field.

For the value of the longitudinal electric field strength, **at a point internal to the conductor** (see Figure I-1), from (9) we have:

$$-\frac{\partial \Phi(x_1,y_1,z,t)}{\partial z}-\frac{\partial A_{mz}(x_1,y_1,z,t)}{\partial t}=E_z(x_1,y_1,z,t). \qquad (1\text{-}1)$$

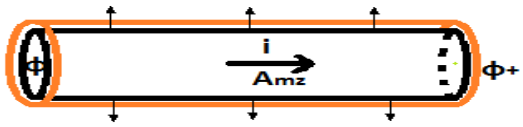


Figure I-1 – Segment of the conductor, of length "Δz", between two scalar potential nodes.

Also, from the equations on the right-hand side of (1) and (3), we have:

$$\nabla\cdot\frac{\partial \vec{B}_e}{\partial t}=\nabla\cdot\left(\vec{J}+\frac{\partial \vec{D}}{\partial t}\right)=\nabla\cdot\vec{J}+\frac{\partial \rho_{efr}}{\partial t}=0. \qquad (1\text{-}2)$$

Integrating (1-2) in a closed surface that involves **a segment of length "Δz" of the horizontal conductor (**see Figure 1-2), we have:

$$\frac{\partial i(z,t)}{\partial z}+I_{LS}(z,t)=-\frac{\partial \lambda(z,t)}{\partial t}. \qquad (1\text{-}3)$$

Where: "i" is the total current flowing through the cross-section of the conductor, "I_{LS}" is the conduction current flowing out of the conductor through the lateral surface, per unit length, and "λ" is the free electric charge accumulated on the surface of the conductor, per unit length.

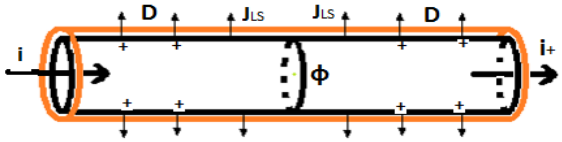


Figure I-2 – Segment of a conductor, of length "Δz", around a scalar potential node.

Equation (1-3) can be interpreted as the application of the Kirchhoff's law to that particular "potential node".

If we separate the potentials produced by the matter existing within the segment "Δz" ("$A_{mz}^{(\Delta)}$" and "$\Phi^{(\Delta)}$"), from the potentials produced by the rest of the matter, we have:

$$\begin{aligned} &A_{mz} = A_{mz}^{(\Delta)} + A_{mz}^{(ext)}; \qquad \Phi = \Phi^{(\Delta)} + \Phi^{(ext)}; \\ &A_{mz}^{(\Delta)}\left(x_1, y_1, z, t\right) = L^* i\left(z,t\right); \quad \lambda\left(z,t\right) = C^* \Phi^{(\Delta)}\left(x_1, y_1, z, t\right). \end{aligned} \tag{1-4}$$

Where: the potentials produced by the rest of the matter, "$A_{mz}^{(ext)}$" and "$\Phi^{(ext)}$", are the potentials produced by the matter existing in the conductor, outside the segment "Δz", plus the potentials "$A_{mz}^{(ex)}$" and "$\Phi^{(ex)}$" representing the externally applied field; and, "L*" and "C*" are, respectively, the inductance per unit length and the capacitance per unit length of the segment "Δz" of the conductor.

Using Ohm's law for the conductor, we have:

$$E_z\left(x_1, y_1, z, t\right) = R_c i\left(z,t\right). \tag{1-5}$$

Where, "R_c" is the resistance, per unit length, of the conductor.

Then, (1-1) and (1-3) can be written as:

$$\frac{\partial \Phi}{\partial z} + R_c i + L^* \frac{\partial i}{\partial t} = -\frac{\partial A_{mz}^{(ext)}\left(x_1, y_1, z, t\right)}{\partial t}; \tag{1-6}$$

and

$$\frac{\partial i}{\partial z} + C^* \frac{\partial \Phi}{\partial t} = C^* \frac{\partial \Phi^{(ext)}\left(x_1, y_1, z, t\right)}{\partial t} - I_{LS}\left(z,t\right). \tag{1-7}$$

As normally, at least a part of the conduction current "I_{LS}", flowing out of the conductor through the lateral surface, is due to the linear leakage current, we can write:

$$I_{LS}\left(z,t\right) = G^* \Phi^{(\Delta)}\left(x_1, y_1, z, t\right) + I_{LS}^*\left(z,t\right). \tag{1-8}$$

And (1-7) can be written as:

$$\frac{\partial i}{\partial z} + G^* \Phi + C^* \frac{\partial \Phi}{\partial t} = G^* \Phi^{(ext)} + C^* \frac{\partial \Phi^{(ext)}}{\partial t} - I_{LS}^*. \tag{1-9}$$

Equations (1-6) and (1-9) have the form of transmission-line equations, and, they are completely general and rigorous equations that describe the interaction of an external electromagnetic field with a straight finite segment of a single-wire line, being the only assumptions in deriving these equations: the thin-wire approximation and the validity of Ohm's law for the conductor.

They can be applied to short horizontal single-wire lines and to long horizontal single-wire lines; also, they can be applied to bent lines, because, in (1-6) and (1-9), "z" represents the direction of the segment of conductor being described, which can be any direction in real space.

Then, they can be applied to different problems, such as, electromagnetic neural stimulation and the calculation of lightning-induced voltages.

As mentioned in reference [1], the main variables in (1-6) and (1-9) are the potentials, which we have seen are very useful for classification purposes. But they are not uniquely determined. Their values, as well as the value of the inductance and capacitance coefficients, are affected by the choice of the reference point, which is the point where the value of the potentials is equal to zero. The reference point is normally chosen at the infinity, and with this choice, from (5) we have, in the Lorenz' gauge, the usual retarded potentials [4]:

$$\begin{aligned} \Phi\left(\vec{r},t\right) &= \frac{1}{4\pi\varepsilon_0} \iiint_{all\ space} \frac{\rho_{et}\left(\vec{r}\,',t\right)}{\left|\vec{r}-\vec{r}\,'\right|} dV'; \\ \vec{A}_m\left(\vec{r},t\right) &= \frac{\mu_0}{4\pi} \iiint_{all\ space} \frac{J_{et}\left(\vec{r}\,',t\right)}{\left|\vec{r}-\vec{r}\,'\right|} dV'. \end{aligned} \tag{1-10}$$

Another usual choice assumes the existence of a line along the z direction, where:

$$E_z\left(ref\right) = A_{mz}\left(ref\right) = \Phi_m\left(ref\right) = 0. \tag{1-11}$$

This line, which is taken as the reference, could be located at the infinite or very far from the structures of interest, in such a way that the potentials produced by the matter behind the reference are negligible.

Sometimes the reference is chosen on a perfectly conducting conductor. When this perfect conductor has an infinite plane surface, the potentials produced by the matter behind the reference, are represented by the potentials produced by the images of the matter on the perfect conductor plane [13].

With any choice of reference, as also mentioned in reference [1], Φ in (1-6) and (1-9), can also be interpreted as the **potential difference** between the conductor and the reference, at a value of z and t.

In the classical transmission-line theory, which is geared for long horizontal transmission lines, the values utilized, for the capacitance per unit length "C" and for the inductance per unit length "L" of the line, are those calculated, neglecting the retardation effects, for an infinite straight line which is uniformly charged with a charge density, per unit length, equal to "λ" and with a current "i" equal in all segments along the line [12]. In this situation, the potentials produced are the so-called *static potentials* "$A_{mz}^{(st)}$" and "$\Phi^{(st)}$".

We will define the so-called *internally-produced potentials* "$A_{mz}^{(int)}$" and "$\Phi^{(int)}$", which are the potentials really produced by the matter existing in the conductor; and, we will also define the potentials "$A_{mz}^{(c)}$" and "$\Phi^{(c)}$", which are the difference between the internally-produced potentials and the static potentials.

$$\begin{aligned} &A_{mz}^{(int)} \equiv A_{mz} - A_{mz}^{(ex)}; \quad A_{mz}^{(c)} \equiv A_{mz}^{(int)} - A_{mz}^{(st)}; \quad A_{mz}^{(st)} = Li\left(z,t\right); \\ &\rightarrow A_{mz} = A_{mz}^{(st)} + \left(A_{mz}^{(ex)} + A_{mz}^{(c)}\right); \\ &\Phi^{(int)} \equiv \Phi - \Phi^{(ex)}; \qquad \Phi^{(c)} \equiv \Phi^{(int)} - \Phi^{(st)}; \quad \lambda\left(z,t\right) = C\Phi^{(st)}; \\ &\rightarrow \Phi^{(st)} = \Phi - \left(\Phi^{(ex)} + \Phi^{(c)}\right); \qquad G^*\Phi^{(\Delta)} = G\Phi^{(st)}. \end{aligned} \tag{1-12}$$

Writing (1-6) and (1-9) in terms of the quantities defined in (1-12), we have:

$$\frac{\partial \Phi}{\partial z} + R_c i + L \frac{\partial i}{\partial t} = -\frac{\partial\left(A_{mz}^{(ex)} + A_{mz}^{(c)}\right)}{\partial t}; \tag{1-13}$$

and

$$\frac{\partial i}{\partial z}+G\Phi+C\frac{\partial \Phi}{\partial t}= \\ =G\left(\Phi^{(ex)}+\Phi^{(c)}\right)+C\frac{\partial\left(\Phi^{(ex)}+\Phi^{(c)}\right)}{\partial t}-I^{*}_{LS}. \tag{1-14}$$

To allow an easy comparison with the equations utilized by the people interested in lightning-induced voltages, we will write (1-13) and (1-14) in terms of the internally-produced scalar potential. Then, we have:

$$\frac{\partial \Phi^{(\text{int})}}{\partial z}+R_c i+L\frac{\partial i}{\partial t}=-\left(\frac{\partial \Phi^{(ex)}}{\partial z}+\frac{\partial A^{(ex)}_{mz}}{\partial t}\right)-\frac{\partial A^{(c)}_{mz}}{\partial t}; \tag{1-15}$$

and

$$\frac{\partial i}{\partial z}+G\Phi^{(\text{int})}+C\frac{\partial \Phi^{(\text{int})}}{\partial t}=C\frac{\partial \Phi^{(c)}}{\partial t}+G\Phi^{(c)}-I^{*}_{LS}. \tag{1-16}$$

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (1-15) and (1-16) can be written as:

$$\frac{d\Phi^{(\text{int})}_{\omega}}{dz}+\left(R_c+i\omega L\right)I_{\omega}=E^{(ex)}_{z\omega}-i\omega A^{(c)}_{mz}; \tag{1-17}$$

and

$$\frac{dI_{\omega}}{dz}+\left(G+i\omega C\right)\Phi^{(\text{int})}_{\omega}=\left(G+i\omega C\right)\Phi^{(c)}_{\omega}-I^{*}_{LS\omega}. \tag{1-18}$$

To show the generality of (1-15) and (1-16), we will compare (1-17) and (1-18) with reference [13]; where the problem of a single-wire line above a perfectly conducting ground, in presence of an impinging electromagnetic field, is considered.

The result there presented is a system of "generalized" or "full-wave" equations, containing electrodynamics corrections (including radiation) to the standard theory of transmission lines. Before making the comparison we must note that:

- In reference [13], the total fields are not divided in: internally produced by the single-wire and externally produced, but, they are divided in what they call "scattered fields", which are produced by both: the single-wire line and its image in the perfectly conducting ground, and what they call "the *exciting electric field* "E^e", which is obtained by the sum of the *incident field* "E^i" and the ground-reflected field "E^r", both determined in the absence of the wire".
- Reference [13] only considers an externally impinging electromagnetic field, neglecting the eventual localized external sources conductively connected to the single-wire. Then, the *internally-produced potentials* reduce to the *scattered potentials*.

Reference [13] calculates the potentials produced by the matter really existing in the conductor, outside the segment "Δz", considering the time delay in the production of the potentials. They in fact calculate the potentials "$A_{mz}^{(c)}$" and "$\Phi^{(c)}$".

Then, when the perfectly conducting ground is taken as the reference for the potentials, neglecting the conductor resistance, per unit length, "R_c" and the transversal currents "I_{LS}", (1-17) and (1-18) reduce to (9) and (10) of reference [13].

2. Comparison of our Field Coupling model with standard coupling theories

Here, as in part 3, we will adopt the same terminology utilized in reference [1].

In order to allow an easy comparison with classical transmission-line theories, we will write (34) and (35) in terms of the internally-produced scalar potential. Then, we have:

$$\frac{\partial \Phi^{j(\text{int})}}{\partial z}+\left(R_{(cj)}i^{j}+\left[L^{j}_{k}\right]\frac{\partial i^{k}}{\partial t}\right)=E^{j(ex)}_{z}-\frac{\partial A^{j(c)}_{mz}}{\partial t}; \tag{2-1}$$

and

$$\frac{\partial i^{j}}{\partial z}+\left[G^{j}_{k}\right]\Phi^{k(\text{int})}+\left[C^{j}_{k}\right]\frac{\partial \Phi^{k(\text{int})}}{\partial t}= \\ =\left[G^{j}_{k}\right]\Phi^{k(c)}+\left[C^{j}_{k}\right]\frac{\partial \Phi^{k(c)}}{\partial t}-I^{j*}_{LS}. \tag{2-2}$$

Also, in order to compare our results with reference [14], which is considered to represent the classical transmission-line theory [13], for this quasi-longitudinal mode (because of the presence of "I_{LS}"), neglecting the transversal magnetic potential "$\mathbf{A}_{mt}$" produced by the transversal current "I_{LS}", we will define the *internally-produced line voltage* with respect to the reference conductor, as in reference [14]:

$$V^{j(\text{int})}(z,t)=\Phi^{j(\text{int})}\left(x_j,y_j,z,t\right)-\Phi^{0(\text{int})}\left(x_0,y_0,z,t\right). \tag{2-3}$$

The *internally-produced line voltage*, defined in (2-3), when the external sources conductively connected to the line are not considered, is identical to the so-called *scattered line voltage*, defined in (7) of reference [14], which is usually adopted in the transmission-line theory utilized by the people interested in lightning-induced voltages [13].

Using the definition of (2-3), (2-1) and (2-2) can be written as:

$$\frac{\partial V^{j(\text{int})}}{\partial z}+\left(R_{(cj)}i^{j}+\left(\left[L^{j}_{k}\right]-\left[L^{0}_{k}\right]\right)\frac{\partial i^{k}}{\partial t}\right)= \\ =E^{j(ex)}_{z}+\left(R_{(c0)}i^{0}-E^{0(ex)}_{z}\right)-\frac{\partial A^{j(c)}_{mz}}{\partial t}+\frac{\partial A^{0(c)}_{mz}}{\partial t}; \tag{2-4}$$

and

$$\frac{\partial i^{j}}{\partial z}+\left[G^{j}_{k}\right]V^{k(\text{int})}+\left[C^{j}_{k}\right]\frac{\partial V^{k(\text{int})}}{\partial t}=-\left[G^{j}_{k}\right]\Phi^{0(\text{int})}+ \\ -\left[C^{j}_{k}\right]\frac{\partial \Phi^{0(\text{int})}}{\partial t}+\left(\left[G^{j}_{k}\right]\Phi^{k(c)}+\left[C^{j}_{k}\right]\frac{\partial \Phi^{k(c)}}{\partial t}-I^{j*}_{LS}\right). \tag{2-5}$$

If we separate, from the current in the reference conductor, the part corresponding to the return current of the other conductors, we can define:

$$i^{0}=i^{0*}-\sum_{k\neq 0}i^{k}. \tag{2-6}$$

Then, (2-4) can be written as:

$$\frac{\partial V^{j(\text{int})}}{\partial z}+\left(\left[R^{j}_{k}\right]i^{k}+\left(\left[L^{j}_{k}\right]-\left[L^{0}_{k}\right]\right)\frac{\partial i^{k}}{\partial t}\right)= \\ =E^{j(ex)}_{z}-E^{0(ex)}_{z}+R_{(c0)}i^{0*}-\frac{\partial}{\partial t}\left(A^{j(c)}_{mz}-A^{0(c)}_{mz}\right); \tag{2-7}$$

To show the generality of (34) and (35), we will compare (2-7) and (2-5), which are equivalent to (34) and (35), with the corresponding equations of reference [14]:
- Neglecting the differences between the static magnetic potentials and the internally-produced magnetic potentials, the unbalanced current in the reference conductor, and if, from the externally applied electromagnetic field, only the externally impinging or incident electromagnetic field is considered, (2-7) reduce to (16) or (33) of reference [14];
- Neglecting the differences between the static magnetic potentials and the internally-produced magnetic potentials (the electrodynamics corrections (including radiation) mentioned in reference [13]), the non-linear conduction current I_{LS}^{*}, flowing out of the "j" conductor through the lateral surface, and assuming that the internally-produced line voltage is equal to the scattered line voltage, (2-5) almost reduce to (32) of reference [14].
The terms remaining in the right-hand side of (2-5) come from an assumption that is made here, which is different from the respective assumption made in reference [14]: The capacitance coefficients are here defined in the Maxwell's way, in terms of the scalar potentials in the conductors; while in reference [14] they are defined in terms of the *scattered line voltages* with respect to the reference conductor.
Then, as the scattered line voltages represent only differential modes, the remaining terms represent the conversion of common mode into differential mode.

When the *internally-produced line voltage* "U^{j}", in the conductor "j", is defined with respect to the reference line for the potentials, we have:

$$U^{j(\text{int})}(z,t) = \Phi^{j(\text{int})}(x_j, y_j, z, t). \tag{2-8}$$

Using the definition of (2-8), (2-1) and (2-2) now read as:

$$\frac{\partial U^{j(\text{int})}}{\partial z} + \left(R_{(cj)} i^{j} + \left[L_k^j \right] \frac{\partial i^k}{\partial t} \right) = E_z^{j(ex)} - \frac{\partial A_{mz}^{j(c)}}{\partial t}; \tag{2-9}$$

and

$$\begin{aligned} &\frac{\partial i^j}{\partial z} + \left[G_k^j \right] U^{k(\text{int})} + \left[C_k^j \right] \frac{\partial U^{k(\text{int})}}{\partial t} = \\ &= \left[G_k^j \right] \Phi^{k(c)} + \left[C_k^j \right] \frac{\partial \Phi^{k(c)}}{\partial t} - I_{LS}^{j*}. \end{aligned} \tag{2-10}$$

Equations (2-9) and (2-10), which are equivalent to (34) and (35), are also completely general and rigorous equations that describe the interaction of an external electromagnetic field with the considered filamentary multi-conductor transmission line, being the only assumptions in deriving these equations: the thin-wire approximation for all the conductors and the validity of Ohm's law for all the conductors.

Now, if we compare (2-9) and (2-10), with the corresponding equations of reference [14]:
- Neglecting the differences between the static magnetic potentials and the internally produced magnetic potentials (the electrodynamics corrections (including radiation) mentioned in reference [13]), and if, from the externally applied electromagnetic field, only the externally impinging or incident electromagnetic field is considered, (2-9) reduce to (16) or (33) of reference [14];
- Neglecting all the terms of the right-hand side of (2-10) that represent electrodynamics corrections (including radiation) and the non-linear transversal conduction current I_{LS}^{*}; and, assuming that the internally-produced line voltage is equal to the scattered line voltage, (2-10) reduce to (32) of reference [14].
Then, we also obtain the advantage for the numerical solution of the system of equations, claimed in reference [14], of having the source term appearing in only one kind of equations.
In fact, this is the formulation utilized for the numerical examples shown in reference [14], which refer to a two-conductor line over a perfectly conducting ground that is taken as a reference plane, not as a reference conductor.

Equations (34) and (35), which are a generalization of (21c) and (22c) of reference [1], as mentioned in reference [1], also represent a generalization of Rusck's coupling theory [21], by including, besides the corrections due to the time delay in the production of the potentials, the effect of the externally produced magnetic potential along the direction of the line (a shortcoming of that theory pointed out a long time ago [39]) and also the effect of the conductive imperfections of the matter, both of the conductors and of the dielectrics between the conductors.

To include the effect of the transversal mode, we will define the so-called *conductor voltages*.
Usually, in standard transmission-line theory, the *conductor voltages* are defined relative to one of the conductors chosen to be the reference [12,30]. Then, to include the effect of the transversal mode, here, we will also define the *conductor* "j" *voltage* as the integral of the transversal electric field strength "$\mathbf{E}_t$" between the conductor "j" and the reference conductor:

$$\begin{aligned} &V^{j(\text{int})} + V^{j(ex)} = V^{j}(z,t) \equiv -\int_{(x_0,y_0)}^{(x_j,y_j)} \vec{E}_t \cdot d\vec{l}_t = \\ &= \Phi^{j} - \Phi^{0} + \int_{(x_0,y_0)}^{(x_j,y_j)} \frac{\partial \vec{A}_{mt}^{(ex)}}{\partial t} \cdot d\vec{l}_t. \end{aligned} \tag{2-11}$$

Where, as usual, neglecting the transversal magnetic potential produced by the transversal currents, "A_{mt}" is the externally produced field.

Then, we can modify (34) and (35), by means of (2-11), and obtain:

$$\begin{aligned} &\frac{\partial V^j}{\partial z} + \left(\left[R_k^j \right] i^k + \left(\left[L_k^j \right] - \left[L_k^0 \right] \right) \frac{\partial i^k}{\partial t} \right) = \\ &= E_z^{j(ex)} - E_z^{0(ex)} + \frac{\partial V^{j(ex)}}{\partial z} + \left(R_{(c0)} i^{0*} - \frac{\partial}{\partial t} \left(A_{mz}^{j(c)} - A_{mz}^{0(c)} \right) \right). \end{aligned} \tag{2-12}$$

and

$$\begin{aligned} &\frac{\partial i^j}{\partial z} + \left[G_k^j \right] V^k + \left[C_k^j \right] \frac{\partial V^k}{\partial t} = \left[G_k^j \right] V^{k(ex)} + \left[C_k^j \right] \frac{\partial V^{k(ex)}}{\partial t} + \\ &+ \left(\left[G_k^j \right] \left(\Phi^{k(c)} - \Phi^{0(\text{int})} \right) + \left[C_k^j \right] \frac{\partial}{\partial t} \left(\Phi^{k(c)} - \Phi^{0(\text{int})} \right) - I_{LS}^{j*} \right) \end{aligned} \tag{2-13}$$

In equations (2-12) and (2-13), which are a generalization of (28) and (29) of reference [1], we can also see, as mentioned in reference [1], that writing the equations in terms of the *conductor voltages* produce new source terms in (2-12) and (2-13), due to the existence of induced voltages produced by the externally applied electromagnetic field.

Apart from the assumptions of the thin-wire approximation and the validity of Ohm's law for all the conductors, and the assumption that the transversal mode is only externally produced, (2-12) and (2-13) are quite general. Neglecting the terms that appear within the parenthesis in the right-hand side of (2-12) and (2-13), they reduce to (3a) and (3b) of reference [12].

For the special case of time-harmonic electromagnetic fields, or in frequency domain, (2-12) and (2-13) can be written as:

$$\frac{dV_{\omega}^{j}(z)}{dz}+\left(\left[R_{k}^{j}\right]+i\omega\left[L_{k}^{j}\right]\right)I_{\omega}^{k}(z)=S_{E\omega}^{j}(z); \qquad (2\text{-}14)$$

and

$$\frac{dI_{\omega}^{j}(z)}{dz}+\left(\left[G_{k}^{j}\right]+i\omega\left[C_{k}^{j}\right]\right)V_{\omega}^{k}(z)=S_{H\omega}^{j}(z). \qquad (2\text{-}15)$$

Which, without the sources, are the standard transmission-line equations in frequency domain, such as shown in (2) of reference [30] or (6.7) and (6.8) of reference [31].

References

[1] P. E. Munhoz-Rojas, "Conducted and Radiated Interference on the interconnections of Renewable Energy Farms" in Proceedings of the 2018 Joint IEEE EMC &APEMC, Singapore, 798-804, 2018.

[2] CIGRE Technical Brochure 719, "Power Quality and EMC issues wih future electricity networks", JWG C4-24/CIRED, www.e-cigre.org, 2018, p.56 .

[3] C. L. Holloway, E.F. Kuester, A.E. Ruehli and G. Antonini, "Partial and internal inductance: Two of Clayton R. Paul's many passions", IEEE Trans. on EMC 55(4), 600-613, 2013.

[4] J.D. Jackson, Classical Electrodynamics, 3rd ed., New York: John Wiley & Sons, Inc., 1999.

[5] J.C. Maxwell, "A dynamical theory of the electromagnetic field", Roy. Soc. Trans. Vol.155, 526-597, 1864.

[6] A. Sommerfeld, Electrodynamics, New York: Academic Press, 1952.

[7] P. E. Munhoz-Rojas, "A Symmetrical Theory of Electromagnetism" in PIERS Proceedings - Moscow2012, 1199-1203, 2012.

[8] R.P. Feynman, R. Leighton and M. Sands, The Feynman Lectures on Physics - The Electromagnetic Field, vol. 2, Reading, Massachusetts: Addison-Wesley, 1965.

[9] Landau, L.D., and E.M. Lifshitz, Course of Theoretical Physics - Vol. 8: Electrodynamics of continuous media, (Spanish translation) Editorial Reverté, Barcelona, 1975, pp. 285-286.

[10] Landau, L.D., and E.M. Lifshitz, Course of Theoretical Physics - Vol. 2: Field Theory, 2nd ed.,(Portuguese translation) Editora Mir, Moscow, 1980, pp. 31, 69.

[11] S.A. Schelkunoff, "Forty years ago: Maxwell's theory invades engineering - and grows with it", IEEE Trans. on Education 15(1), 2-14, 1972.

[12] C. R. Paul, "A brief history of work in transmission lines for EMC applications", IEEE Trans. on EMC 49(2), 237-252, 2007.

[13] F. Rachidi, "A review of field-to-transmission line coupling models with special emphasis to lightning-induced voltages on overhead lines", IEEE Trans. on EMC 54(4), 898-911, 2012.

[14] A.K. Agrawal, H.J. Price and S.H. Gurbaxani, "Transient response of multiconductor transmission lines excited by a nonuniform electromagnetic field", IEEE Trans. on EMC 22(2), 119-129, 1980.

[15] P. E. Munhoz-Rojas, "The effect of discontinuities in a multiconductor line on lightning-induced voltages", IEEE Trans. on EMC 51(1), 53-66, 2009.

[16] A. L. Hodgkin and A. F. Huxley, "A quantitative description of mem-brane current and its application to conduction and excitation in nerve," J. Physiol., vol. 117, 500-544, 1952.

[17] D. R. McNeal, "Analysis of a model for excitation of myelinated nerve," IEEE Trans. Biomed. Eng., vol. BME-23, 329-337, 1976.

[18] J. P. Reilly, V. T. Freeman and W. D. Larkin, "Sensory effects of transient electrical stimulation – Evaluation with a neuroelectric model" IEEE Trans. Biomed. Eng., vol. BME-32, 1001-1011, 1985.

[19] F. Rattay, "Analysis of models for external stimulation of axons" IEEE Trans. Biomed. Eng., vol. BME-33, 974-977, 1986.

[20] B. J. Roth and P. J. Basser, "A model of the stimulation of a nerve fiber by electromagnetic induction" IEEE Trans. Biomed. Eng., vol. BME-37, 588-597, 1990.

[21] S. Rusck, "Induced lightning overvoltages on power transmission lines with special reference to the overvoltage protection of low voltage networks", in *Trans. of the Royal Institute of Technology*, Stockholm, Sweden, vol.120, 1958.

[22] T.W. Dawson, J. de Moerloose and M.A. Stuchly, "Hybrid finite-difference method for high-resolution modelling of low-frequency electric induction in humans", J. of Comp. Phys. 136, 640-653, 1997.

[23] X.L. Chen, S. Benkler, C.H. Li, N. Chavannes and N. Kuster, "Low frequency electromagnetic field exposure study with posable human body model" in 2010 IEEE International Symposium on Electromagnetic Compatibility, Fort Lauderdale, FL, USA, 702-705, 2010.

[24] P. E. Munhoz-Rojas, C.S. Segura-Salas, A.A. Costa, R. Martins and J. Hoffmann, "Fields and current densities induced in the human body by low-frequency electromagnetic fields" in Proceedings of the 2018 Joint IEEE EMC &APEMC, Singapore, 1267-1273, 2018.

[25] J.C.Maxwell, A Treatise on Electricity and Magnetism, 2nd Edition, Vol. II, Oxford: Clarendon Press, 1881, pp.207-208.

[26] J.C.Maxwell, A Treatise on Electricity and Magnetism, Vol. I, Oxford: Clarendon Press, 1873, pp.89-90.

[27] L. F. Shampine, Numerical Solution of Ordinary Differential Equations. New York: Chapman & Hall, 1994.

[28] C. R. Paul and J. W. McKnight, "Prediction of crosstalk involving twisted pairs of wires-Part I: A Transmission-line model for twisted-wire pairs", IEEE Trans. on EMC 21(2), 92-105, 1979.

[29] G. Spadacini, F. Grassi, F. Marliani and S. A. Pignari, "Transmission-line model for field-to-wire coupling in bundles of twisted-wire pairs above ground", IEEE Trans. on EMC 56(6), 1682-1690, 2014.

[30] S. Greedy, C. Smartt, M, Basford and D. Thomas, "Open source cable models for EMI simulations", IEEE EMC Magazine 7(3), 69-81, 2018.

[31] CIGRE Technical Brochure 707, "EMC in wind energy systems", WG C4-30, www.e-cigre.org, 2017, pp.45-46 .

[32] C. A. Nucci, F. Rachidi, M. V. Ianoz and C. Mazzetti, "Lightning-induced voltages on overhead lines", IEEE Trans. on EMC 35(1), 75-86, 1993.

[33] A. Tatematsu, F. Rachidi and M. Rubinstein, "A technique for calculating voltages induced on twisted-wire pairs using the FDTD method", IEEE Trans. on EMC 59(1), 301-304, 2017.

[34] P. Munhoz-Rojas and C. L. da S. Pinto, "Analysis of lightning-induced voltages on an overhead line" in Proc. of. 8th Int. Symp. Lightning Protection, Sao Paulo, Brazil, 57–62, 2005.

[35] S. Yokoyama, "Calculation of lightning-induced voltages on overhead multiconductor system", IEEE Trans. Power App. Syst., vol. PAS-103, no. 1, pp. 100–108, Jan. 1984.

[36] M. Paolone, C. A. Nucci, E. Petrache, and F. Rachidi, "Mitigation of lightning-induced overvoltages in medium voltage distribution lines by means of periodical grounding of shielding wires and of surge arresters: Modeling and experimental validation", IEEE Trans. Power Del., 19(1), 423–431, 2004.

[37] P. E. Munhoz-Rojas and C.L.da S, Pinto, "Calculations of lightning-induced voltages in distribution lines with LSA", Journal of Energy, 60, Special Issue, 95-100, 2011.

[38] Hertz H., "The forces of electric oscillations, treated according to Maxwell's theory" (in German), *Wiedemann's Annalen*, 36, p.1, 1889. (English translation in Electric Waves, by Heinrich Hertz, translated by D.E. Jones, MacMillan and Co., London and New York, IX,137-159, 1893).

[39] V. Cooray, "Calculating lightning-induced overvoltages in power lines: A comparison of two coupling models", IEEE Trans. on EMC 36(3), 179-182, 1994.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 3, 145-149 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Transfer Function Analysis of Fractional-Order Three-Dimensional Electrically Coupled Cell Network

Mahmut ÜN*

Department of Electrical-Electronics Engineering, Faculty of Engineering and Architecture, İstanbul Yeni Yüzyıl University, Istanbul 34010, Turkey, mahmut.un@yeniyuzyil.edu.tr

A R T I C L E I N F O

Article history:
Received: 13 December, 2018
Accepted: 19 May, 2019
Online: 10 June, 2019

Keywords:
Fractional-order
Transfer function analysis
Impedance characteristics
Cell network
Comparative analysis

A B S T R A C T

In this paper, a novel method is proposed for the dynamic analysis of fractional- order three-dimensional electrically coupled cell network. In general, three- dimensional cell network is constructed by combining three one-dimensional circuit networks. Analysis method is based on the principles of the dynamic analysis with transfer function approximation. Although fractional-order three-dimensional circuit network model contains nonlinear fractal elements such as fractional-order capacitors and inductors, transfer function approximation is employed in dynamic analysis of the network by using the Laplace transform.

First by using nodal analysis method, general expression in matrix forms for the transfer function and typical equivalent impedance of the fractional-order three-dimensional circuit network are derived in fractional domain. Transfer function and Equivalent network impedance of the cell network model are obtained in the form of matrix equation with the implicit analytical expression. Secondly the effects of five network parameters such as inductance L, capacitance C, number of cell unit n and fractional- orders (α, β) on the impedance and electrical network characteristics such as transfer function and output responses are investigated by means of MATLAB Simulation programs. Finally, the validity of the proposed method is done by using PSPICE simulations which show the experimental performance and PSPICE simulation results is presented.

1. Introduction

Due to the use of electrical and non-electrical systems for modeling, in recent years circuit networks have become more attractive [1,2]. A study on the resistance network of the grapyhene showing the presence of a planar circuit network in nature received the Nobel Prize in Physics in 2010 [3-5]. In the last few decades, researchers have published publications on fixed-coefficient circuit networks. These publications focus on the analysis of circuit networks consisting of single-member resistors and capacities with constant coefficients [6,7].

The fractional order mathematical models developed for inductances and capacities have more accurate electrical characteristics for the better realization and best fit. Especially actual inductors and capacitors are in fractional order elements in nature [8]. In recent years, some researchers have been working on the realization and design of fractional elements [9-11]. In addition, in recent years, researchers have focused on fractional order theory [12 -15]. However, few researchers have dealt with the issue of electrical characteristics of multidimensional circuit networks. Therefore, we focus on the equivalent network impedance and transfer function of the three-dimensional cell networks in the fractional domain.

Recently, some researches concentrate on the fractional-order three-dimensional circuit network [16]. Moreover, there has been little research on the transfer function and input equivalent impedance characteristics of the three-dimensional circuit network in the fractional order sense with different cases. That is why, we concentrate on the electrical characteristics of the fractional-order three- dimension network, which may lay the bases for both the electric circuits and dielectrics communities [17].

*Mahmut ÜN, Electrical-Electronic Engineering Department, İstanbul Yeni Yüzyıl University, 34010 İstanbul, Turkey Email: mahmut.un@yeniyuzyil.edu.tr

This study is organized as follows: Basic definitions of the fractional order capacitors and inductors are introduced in Section 2. Also the general expressions of the transfer function and equivalent input impedances of the circuit-network are derived by using the nodal analysis method. MATLAB and PSPICE simulations of the fractional order three- dimensional network are done and simulation results is given in Section 3. Finally the paper is concluded in the last section.

2. Fractional-Order Three Dimensional Circuit Network Model

Electrical networks consisting of one-dimensional and two-dimensional RC elements mentioned in the literature have been proposed and used. Since there are no fractional elements in these electrical networks, the models have been insufficient to precisely characterize the system. RLC elements are used in electrical equivalent circuits in one-dimensional and two-dimensional networks used in the dynamic analysis of electrical interactive cells [18]. Three dimensional random RC networks have been proposed for modeling electrical circuits and dielectric material communities [19]. Nodal-matrix analysis was used to model the three-dimensional network of dispersed capacitance and resistance elements for the multiple bound layers [20]. This article shows that large three-dimensional RC networks can be modeled in terms of performance using fractional order models and integrals. In the fractional domain, the fractional order 2 × n RLC circuit network model has been proposed to examine the impedance characteristics of the circuit network [21]. Electrical RC equivalent circuits are proposed to examine the behavior of the human body impedance against contact currents in a wide frequency range. With the proposed fractional-order three-dimensional RLC networks, expressions for calculating the typical equivalent impedances of the network using the differential equation model and the matrix transform method for different cases in the fractional domain are derived [22].

The circuit diagram of fractional-order three-dimensional electric circuit network used in this study is given in Figure 1.

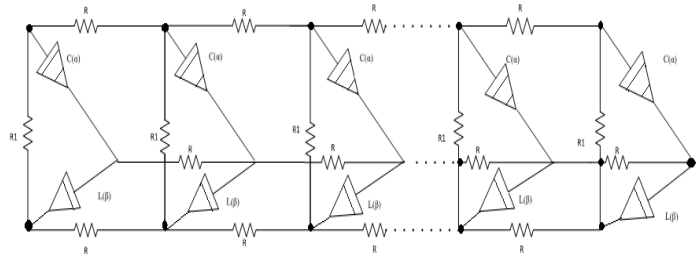


Figure 1: RLC circuit network diagram

In general, this circuit network is obtained by combining three one-dimensional circuit networks. There are n cells in the network. The electrical coupling between the cells is indicated by the resistances R. The cells were modeled with resistance (R_1), fractional- order inductance (L_β) and capacities ($C\alpha$). Fractional impedance values of capacitance and inductance are given below respectively.

$$Z(C_\alpha) = 1 / (C\, s^\alpha) \quad (1)$$

$$Z(L_\beta) = L\, s^\beta \quad (2)$$

Fractional-order three-dimensional cell network is modelled with lumped parameters passive elements such as resistance, fractional-order inductance and capacitance. Thus, this network is composed of passive elements and is a passive network.

In calculating the characteristic values of the circuit network, such as transfer function and the equivalent impedance, the well-known node voltages analysis in circuit theory is used in the fractional domain. For the circuit network depicted in Fig. 1, the equation of the node voltages in matrix form in the fractional domain is written as

$$\begin{bmatrix} Y & -IG & \underline{0} & \underline{0} & \cdots & \underline{0} & \underline{0} \\ -IG & Y_1 & \underline{0} & \underline{0} & \cdots & \underline{0} & \underline{0} \\ & & \vdots & & & & \\ & & \vdots & & & & \\ & & \vdots & & & & \\ \underline{0} & \underline{0} & \underline{0} & \cdots & -IG & Y_1 & -IG \\ \underline{0} & \underline{0} & \underline{0} & \cdots & \underline{0} & -IG & Y \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ \vdots \\ \vdots \\ V_{n-1} \\ V_n \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ \vdots \\ \vdots \\ I_{n-1} \\ I_n \end{bmatrix} \quad (3)$$

Here, the I_1 (3 × 1) indicates the source currents connected to the nodes and the Vi (3 × 1) indicates the node voltages. Vi = $[V_a, V_b, V_c]^t$ shows the node voltages for each cell, where a,b,c nodes are terminals of abc triangle, I (3 × 3) unit matrix, G = 1 / R is resistance conductivity. Elements of the node admittance matrix are;

$$Y = \begin{bmatrix} s^\alpha C + G + G_1 & -G_1 & -s^\alpha C \\ -G_1 & \frac{1}{s^\beta L} + 2G + G_1 & -\frac{1}{s^\beta L} \\ -s^\alpha C & -\frac{1}{s^\beta L} & s^\alpha C + \frac{1}{s^\beta L} + G \end{bmatrix} \quad (4)$$

$$Y_1 = \begin{bmatrix} \frac{1}{s^\alpha C} + G + G_1 & -G_1 & -\frac{1}{s^\alpha C} \\ -G_1 & s^\beta L + 2G + G_1 & -s^\beta L \\ -\frac{1}{s^\alpha C} & -s^\beta L & s^\beta L + \frac{1}{s^\alpha C} + 2G \end{bmatrix} \quad (5)$$

From (3) and (4) and (5) ones can obtain (6) in the matrix form.

$$AV_d = I \quad (6)$$

The node A (3n × 3n) in size is the admittance matrix and is in the form of a symmetric band matrix. It illustrates the vector of the V_d (3n × 1) node voltages vector and I (3n × 1) dimension of the node voltages vector and I (3n × 1) the current vector. If the vector of the node voltages is resolved from the equation (6),

$$V_d = A^{-1} I \quad (7)$$

is obtained. When a single source is applied to the network, the equivalent driving point impedances and transfer functions for various situations can be found from equation (7) by using the basic definition.

3. Computer Simulations

The element and parameter values selected for the circuit network shown in Figure 1; are given as follows. The cell number is taken as n=5 in all computer simulations. The element values are chosen as R_1= 1Ω, R= 0.1 Ω, L= 1 H and C = 1 F.

3.1. MATLAB Simulations

A constant current source I_1 is applied to the Va node and is taken as Va = V_1. If V_1 input voltage is written in terms of node voltages, V_1= [100 000 000 000 000].V_d . The equivalent impedance is calculated as Z = V_1 The equivalent impedance is calculated as Z = V_1. For fractional order terms, the second order approach was obtained using the value of $s^{0.5s} = (5s^2 + 10s + 1) / (s^2 + 10s + 5)$, Z equivalent input impedance of the network was obtained as a rational function in s [23]. The frequency response, step response and impulse response are obtained from Z network impedance function obtained as a rational function. Then MATLAB simulation results for these rational functions are shown in Figure 2, Figure 3, Figure 4, respectively.

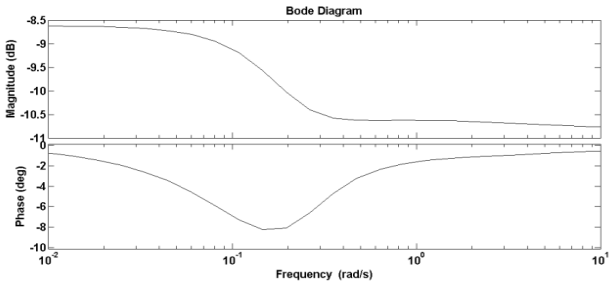


Figure 2: Frequency responses for network input impedance

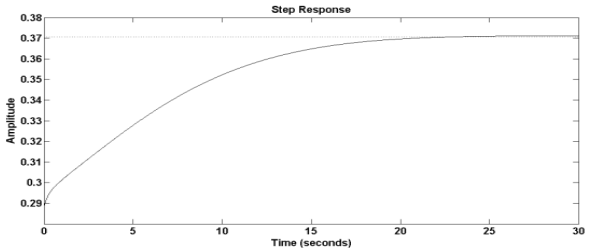


Figure 3*:* Step response of the input impedance

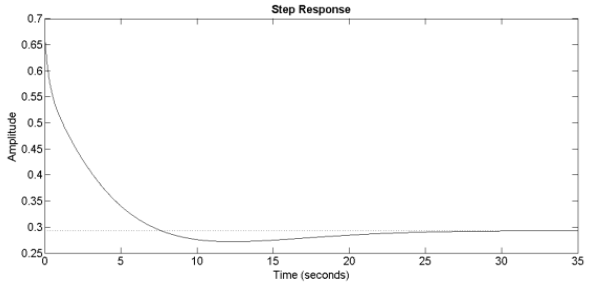


Figure 4*:* Impulse response of the network impedance

Let the output voltage V_o get the input voltage V_1 when the circuit network is excited by a source. The input voltage is V_1= [100 000 000 000 000] .Vd and the output voltage is V_o= [000 000 000 000 001]. Vd. The transfer function (as voltage gain) is defined as TF = V_o / V_1. For fractional order terms, TF is also obtained as a rational function of the TF using the quadratic fractional approach in [24]. The frequency response, step response and impulse response obtained by MATLAB simulations of TF are depicted in Figure 5, Figure 6, Figure 7, respectively.

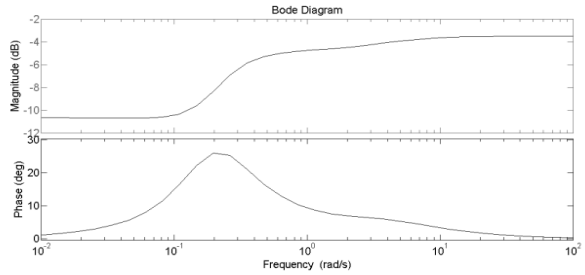


Figure 5*:* The transfer function frequency responses

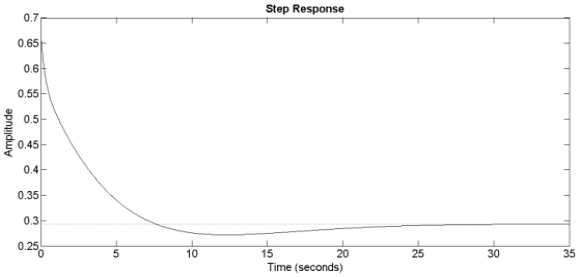


Figure 6*:* The transfer function step response

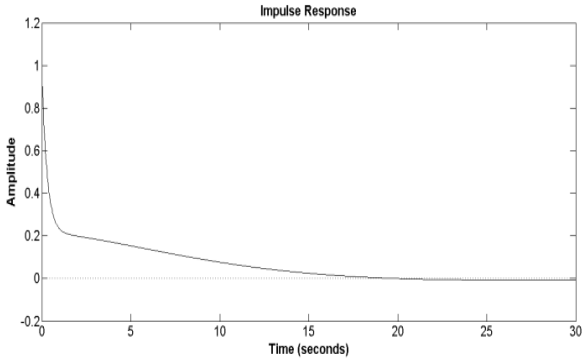


Figure 7*:* The transfer function impulse response

3.2. PSPICE Simulations

The experimental realization of the study was done by PSPICE program using element values taken before in the MATLAB simulations. Developed SPICE model of the network circuit for n=5 cell is presented in Figure 8.

In PSPICE simulations, the quadratic approximation equivalent circuits given in [24] are used instead of fractional capacitor and inductance elements, and the simulation results are depicted in Figure 9, Figure 10, Figure 11, Figure 12, respectively.

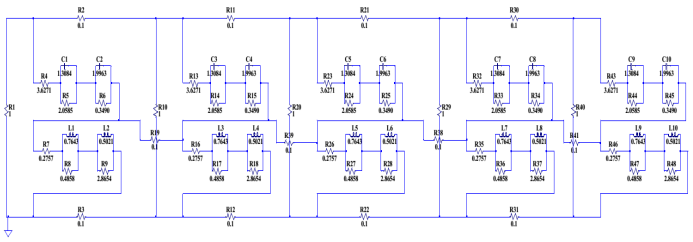


Figure 8: PSPICE model of the cellular network for n=5 cell

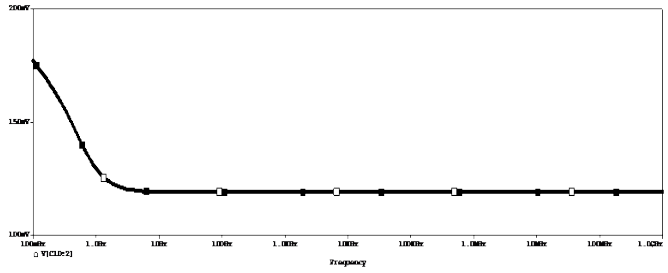


Figure 9: Frequency response of network transfer function with PSPICE

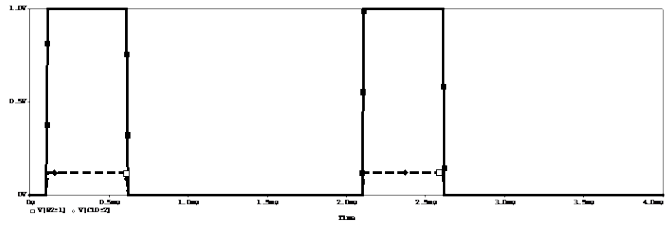


Figure 10: Step response of network transfer function with PSPICE

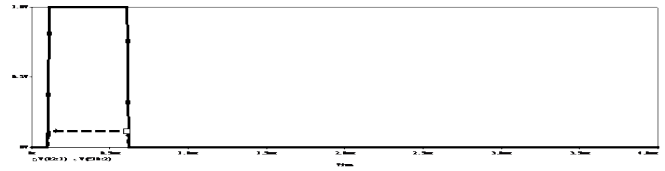


Figure 11: Impulse response for the network transfer function with PSPICE

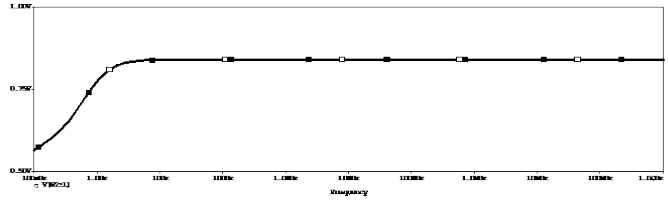


Figure 12: Frequency response of equivalent network impedance with PSPICE

4. Conclusions

In this paper, mathematical models were introduced for the analysis of the fractional-order three-dimensional network in fractional domain. The transfer function and the equivalent input impedance of the fractional-order three-dimensional network were obtained in form of a rational transfer function using rational function approximation for each fractional term. Although the circuit network includes nonlinear fractal elements such as fractional-order capacitors and inductors transfer function approximation method was offered by means of some advanced properties of Laplace transform. A new mathematical method was developed to model and analyze fractional- order three-dimensional circuit network in the fractional domain.

In general, the electrical characteristics of the three-dimensional cellular circuit network were investigated in the fractional domain. Using the nodal analysis method which is well known in circuit theory first transfer function and equivalent input impedance of the fractional order three dimensional circuit network were derived in fractional domain. After that employing second order approach for each fractional order term, equivalent input impedance and transfer function were found as a rational function in s. As a consequence of this process we could be able to calculate transfer function and equivalent input impedance of the network as a rational transfer function.

The electrical system characteristics of the transfer function and the driving point impedance of the fractional- order three–dimensional circuit network were investigated by using MATLAB simulation programs in detail. Moreover network system characteristics such as impulse, frequency, and step responses were graphically obtained for both transfer function and equivalent input impedance. Dynamic characteristics and stability of fractional –order three –dimensional circuit network were also examined with respect to model parameters such as inductance, capacitance, resistances, fractional order values and cell numbers. In order to compare MATLAB and PSPICE simulations the same network parameter values were employed in all simulations. Finally, when PSPICE and MATLAB simulation results are compared graphically, basic similarities between PSPICE and MATLAB simulation results were observed.

References

[1] Von Busse V., Swartz S.M., and C.C, ''Flight metobolism in relation to speed in Chiroptera: Testing the U-shape paradigm in short-tailed fruit bat Carollia perspicillata,'' *J.Exp.Bio.*, 216(11), s. 2073-2080, 2013.

[2] Simpson R., Jaques A., Nunez, C.H. Ramirez, and Almonacid A., ''Fractional calculus as a mathematical tool to improve the modelling of mass transfer phenomena in food,'' *Food Eng.Rev.*, 5(1) s. 45-55, 2013.

[3] Bolotin K.I., Ghahari F., Shulman M.D., Stormer H.L., and Kim p., ''Observation of fractional quantum Hall effect in graphene ,'' *Nature*, 475(7), s. 7354-7358, 2011.

[4] Feldman B.E., Krauss B., Smet J.H., and Yacoby A., "Unconventional sequence of fractıonal quantum hall states in suspended graphene,'' *Science*, 337 (6099), s. 1196-1199, 2012.

[5] Owaidat M.Q., Hijjawi R.S., and Khalifeh J.M., "Network with two extra interstitial resistor," *Int.J.Theor. Phys.*, vol. 10, s. 3152-3159, 2012.

[6] Yao N.Y., Gorshkov A.V., Laumann C.R., Lauchli , J. Ye, and Lukin M.D., "Realizing fractıonal Chern insulators in dipolar spin systems,'' *Phys .Rev.Lett.*, 110(18), 2013.

[7] Hijjawi R.S., Assad J.H., and Sakaji, H.J., "Perturbation of an infinite network of identical capacitors,'' *Int.J. Mod. Phys.*, vol. B2, s. 199-209, 2007.

[8] Carlson G., and Halijaj C., "Approximation of fractional capacitors$(1/s)^{1/n}$ by regular Newton process,'' *IEEE T. Circuit System*, s. 210-213, 1964.

[9] Tenreiro M.J.A., and Galhano A.M.S.F., "Fractional order inductive phenomena based on the skin effect,'' *Nonlinear Dyn.*, vol. 1-2, s. 107115, 2012.

[10] Radwan A.G., and Salama K.N., "Passive and active elements using fractional Lβ,Cα,'' *IEEE T. Circuits-1*, vol. 10, s. 2388-2397, 2011.

[11] Das S., Sivaramakrishna M., Biswas K., and Goswami B., "Performance study of a constant phase angle based impedance sensor to detect milk adultcration. Sensors and actuators,'' *A-Physical*, vol. 2, s. 73-278, 2011.
[12] Krishna M.S., Biswas K., and Goswami B., "Fabrication of a fractional order capacitor with desired specifications; A study on process identification and characterization,'' *IEEE Trans.Electron Dev.*, vol. 11, s. 4067-4073, 2011.
[13] Haba T., Ablart G., Camps T., Olivie F., "Influence of the electrical parameters on the input impedance of a fractal structure reaslised on silicon,'' *Chaos Solitons Fract.*, vol. 2, s. 479, 2005.
[14] Elshurafa A.M., Almadhoun M.N., Salama K.N., and Alshareet H.N., "Microscale electrostatic fractional capacitors using reduced graphene oxide percolated polymer compasites,'' *Appl.Phys.Lett.*, vol. 23, s. 1234-1237, 2013.
[15] Radwan A.G., and Salama K.N., "Fractional–order RC and RL circuit,'' *Circuits System Signal Process*, vol. 6, s. 1901-19015, 2012.
[16] A. Doghri, T.Djerafi, A. Ghotto, and K.Wu, "Substrate integrated wave guide directional couplers for compact three-dimentional integrated circuitsş" IEEE Trans. Microw. Theory Tech., Vol.63, No.1, pp. 209-219, Jan. 2015
[17] R, K, H. Galvao, S. Hadjloucas, K.H. Kienitz, H. M. Paiva, and R.J. M: Afonso,"Fractional order modeling of large three-dimentional RC networks," IEEE Trans. Circuits Syst. I, Reg. Papers, Vol.60, no.3, pp.624-637 , 2013
[18] Chen P., and He S.B., "Analysis of the fractional order paralle tank circuit,'' *J.Circuit System. Comput.*, vo. 6, s. 452-458, 2013.
[19] Merichel M.B., Eyman E.D., and Katar S.B., "Analysis of a Network of Electrically coupled neorans producing Rhytmic Activity in the snail helisoma trivolvis," *IEEE Trans on Biomedical Engineering,* vol. BMM-24, no. 3, s. 277-287, May.1977.
[20] V.Torre, W.G. Owen, G. Sandini, "The Dynamics of electrically interacting cell," *IEEE Trans on Systems, Man and Cybernetics*, vol. 13, no. 5, s. 757-765, September/October 1983.
[21] R. Zhou, D. Chen , Herbert H., C. Lu, "Fractional order 2×n RLC Circuit Network," *Journal.f Circuit Systems and Computers*, vol. 24, no. 9, DOI.10.1142/5021812661550142X(25 pages), 2015.
[22] T.J. Freeborn, A.S. Elwakil, B. Macendy, "Compact wide frequency range fractional order models of human body impedance against contact current," *Hindawi Publishing Corporation, Mathcmatical Problems in Engineering*, vol. 2016, Article ID. 4567937.
[23] R. Tanudar, S. Kumar, "Analysis &Design of Fractance based fractional order filter," Int Journal of Innovative Research in Electrical," *Electronics, Instrumentation and Control Engineering*, vol. 1, Issue. 3, June 2013.
[24] Radwan A. G. And Salama K. N.,"Fractional-order RC and RL circuit", Circuits Sys. Signal Process, Vol:6, 1901-1915, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 153-161 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Frequency-Based Design of Electric System for Off-shore Wind Power Plant (OWPP)

Uttam S. Satpute[*,1], Diwakar R. Joshi[2], Shruti Gunaga[3]

[1]*Department of Electrical and Electronics Engineering, VDIT Haliyal, Karnataka, 581329, India*

[2]*Department of Electrical and Electronics Engineering, GIT Belagavi, Karnataka, 590008, India*

[3]*Department of Electrical and Electronics Engineering, Oxford College of Engineering, Bangalore 560068, India*

ARTICLE INFO

Article history:
Received: 08 January, 2019
Accepted: 25 February, 2019
Online: 20 March, 2019

Keywords:
Optimum Frequency
Offshore wind power plant
Optimization

ABSTRACT

The paper presents a novel concept of design of electric system for Off-shore Wind Power Plant (OWPP) based on frequency. In the literature, a new transmission system i.e. Fractional Frequency Transmission System (FFTS) has been proposed as an economic alternative to High Voltage AC (HVAC) and High Voltage DC (HVDC) for power transmission from Off-shore Wind Farm (OWF) to on-shore grid. In the FFTS power transmission is proposed at 1/3rd of conventional frequency (50Hz). The choice of this frequency (50/3Hz) for FFTS is not based on any mathematical analysis and seems to be a compromise. Since, frequency is one of the prominent parameter that reflects on the project economy, the paper explores on the choice of operational frequency for electric system in view of reduction in the investment cost. The existence of optimum frequency is realized based on the finding that, the costs of power system components are frequency-dependent and they imbibe the trade-off in their costs. The paper presents a comprehensive methodology to compute the most economical operational frequency for electric system by optimization of frequency-based cost model of the system using Genetic Algorithm (GA). The proposed methodology is applied to 160MW OWF (Horns-Rev) as a case study. For off-shore distances of 50km, 100km, 150km and 200km, the optimum operational frequencies are obtained as 52Hz, 32Hz, 22Hz and 16Hz respectively. The proposed methodology for the design of electric system based on frequency results in significant saving in the investment cost for moderate transmission distance. Hence it is concluded that, the work presented in this paper gives a new dimension for planning and cost effective design of electric system for OWPP. Further, it promotes the investor for deployment of OWF which leads to reduction in air pollution and global warming which are major concerns of the world today in the endeavor.

1. Introduction

This paper is an extension of original paper entitled, "Optimum Frequency-Based Design of Off-shore Wind Power Plant (OWPP)", presented in 2017 IEEE PES Asia Pacific Power and Energy Engineering Conference (APPEEC), India [1]. In the present paper, the author has made an attempt to address all the comments, queries and issues raised during the publication of the previous original paper. To avoid de-linking, a brief introduction to the theme of the paper has been presented from the original work.

Harvesting energy from wind has become one of the major contributions to the energy sector across the globe. Due to policy instability as well as specific issues linked to land acquisition for wind power projects on land; deployment of wind project offshore has been increased in the recent times. High-stable offshore wind potential, large installation space, no visual effects, closeness of major load centers and technological advancement in the wind industry has supplemented its growth. Globally offshore installation has been progressed with an average growth rate of 30% per annum during 2010-2017, reaching to a total installed capacity of 18GW at the end of 2017 [2]. Opportunities for further scaling up are humongous. An ambitious target of 40GW by 2020

* Uttam S. Satpute, Email: uttamsatpute@gmail.com

has been reported.

Huge initial investment is one of prominent slog in the path of offshore development. Highly evolved tower construction, offshore substation, seabed lying of submarine cable and high maintenance cost are some of reasons for heavy investment. And it is the major hindrance for offshore deployment. Presently, the Levelized Cost of Energy (LCOE) is more than $5-6 millions per MWh, and it warrants for the technological changes to reduce the investment, and make OWPP economically feasible [3].

In the OWPP, among the split-up, electrical infrastructure cost shares significant portion in the total project cost. It includes the cost of inter-farm distribution network which collects the power from the wind turbines and the cost of main transmission network which exports the power from the off-shore substation to on-shore grid. In broader perceptive, the paper exploits on opportunities to reduce the electrical infrastructure cost for OWPP.

In respect of this, a new transmission system i.e. Fractional Frequency Transmission System (FFTS) has been proposed in the literature as an economic alternative to HVAC and HVDC [4]. FFTS is also named as Low Frequency AC (LFAC) by some researchers. In the LFAC, power transmission is proposed at 1/3rd of the nominal frequency (50/60Hz), and at the grid-side, frequency converter is employed to transform fractional frequency (50/3Hz) to nominal frequency (50Hz). In the LFAC, due to low frequency operation charging current requirement of the cable is less and that increases its power transmission capacity, and unlike HVDC it requires only one converter at the grid-side are main attributes for its cost reduction [4].

The technical and economic viability of LFAC over HVDC and conventional HVAC has been investigated by many researchers [4]-[9]. It is stated that, LFAC has economic and performance advantages over conventional HVAC and HVDC and it has a future trend.

The fundamental quest on the choice 50/3 Hz frequency for LFAC has initiated this paper. Literature states that, single-phase 50/3Hz system is already in-use for supplying electric power to traction in German and in LFAC cyclo-converter can be employed for frequency transformation which is less costly compared to other converters [9]. After a careful perusal of the literature, it is revealed that, the choice of 50/3Hz frequency for LFAC seems to be a compromise, and it is not based on any rigorous mathematical analysis. The choice of frequency for power transmission is very important, because it reflects on the project economy. The original paper has dealt with the choice of operational frequency for electric system integrating OWF to on-shore grid, in view of reduction in the investment cost. Thus, a novel thought of optimum frequency-based design of OWPP had been proposed in the previous work.

In the original work, the reviewers and the panel of judges had raised the following major issues,

The first major issue was related to the methodology adopted for the computation of optimum frequency. In the previous work, optimum operational frequency was computed through optimization of the frequency-based cost model of the electric system. Genetic Algorithm (GA) was used for the optimization. Since the cost model was of single-variable (frequency), the reviewer raised the question on the applicability of GA for the optimization of a simple-single-variable model as it is applied for the case of complex system. In the present work, operating voltage of the electric system is added as another variable for enhancing optimization of the system. Furthermore, in the previous work, cost model of the system had been developed by adding frequency-based cost models of various power system components involved in the system. These cost models of the components were developed separately in the sequence. Firstly the components were designed and then from the design parameters the costs were computed for different frequencies and then from the frequency-based cost curves their cost models were derived using curve fitting techniques. The reviewer suggested combining these processes into a single execution, which is incorporated in the present work. The paper presents a newer computational structure for the computation of the optimum frequency in a single execution.

The second major issue was related to validation of the designs and cost models of the components. Design and frequency-based cost model of the transformer developed in the original paper has been validated with the models available in the literature.

The third major issue was in respect of generation of fractional frequency from the wind turbine generator. In the original work, optimization was performed on Fractional Frequency Wind Power System (FFWPS). Reference [10] has proposed this configuration for integrating OWF to on-shore grid. It eliminates individual power converter in the wind turbine and proposes speed control from single large power converter (cyclo-converter). Since the generator required in this configuration would be bulky as it has to generate low frequency power, it demands for modification in the wind turbine generator. The reviewer had questioned the implications on the design and verification with wind turbine manufacturers. Reference [10] has validated FFWPS for its practical implementation because it reduces the cost of wind turbine generator in twofold; firstly it reduces the gear ratio and secondly it eliminates the power converter. So, in the present paper, FFWPS is considered for the optimization process.

The paper is organized as follows. The structure of FFWPS and the concept of optimum operational frequency are presented in Section 2. Section 3 presents the development of frequency-based cost model of FFWPS. Section 4 presents the development of computation structure for the optimization of the developed model. Section 5 presents the rating of components of FFWPS for application of the developed methodology. The results are discussed in the Section 6 and finally Section 7 concludes the paper.

2. System model (FFWPS) and the concept of optimum frequency

Fractional Frequency Wind Power System (FFWPS) is one of the electric system configuration proposed in the literature for integrating OWF to on-shore grid [10]. Figure 1 shows the structure of this system.

In FFWPS, the entire electric system including generator, inter-farm distribution network and main power transmission network operates at 50/3Hz frequency. Frequency converter at the grid-side transforms frequency from 50/3Hz to 50Hz and the

power is injected onto the grid. FFWPS eliminates the full scale power converter in the wind turbine and proposes rotor speed control from the single large power converter (cycloconverter). Due to elimination of individual power converter the cost is reduced but there is a 3% loss in the energy yield [11]. Reference [13] has discussed synchronization scheme for FFWPS and shown safe and reliable integration of offshore wind farm to onshore grid. From the literature it is revealed that, FFWPS is technically and economically suitable configuration to transmit electrical power from off-shore wind farms to the power grid.

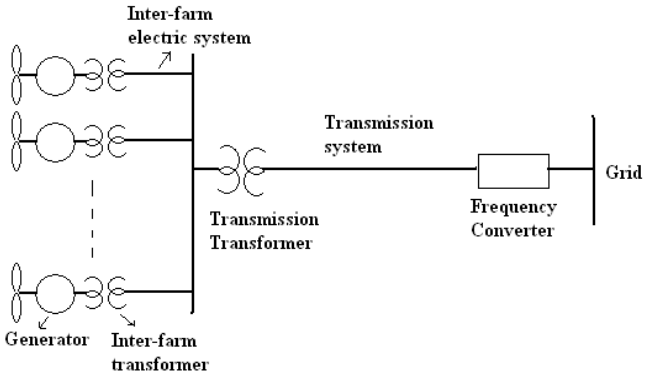


Figure 1: Structure of Fractional Frequency Wind Power System (FFWPS)

In this paper, the study on FFWPS is extended from the perspective of its operational frequency. The operational frequency of FFWPS is 50/3 Hz, but as discussed in the previous section, the choice of 50/3 Hz frequency for FFWPS seems to be compromise rather than based on any formal mathematical analysis. The paper particularly, deals with the choice of frequency (economically-optimal) for FFWPS based on the criterion of minimum investment cost.

The existence of optimum frequency for FFWPS is realized based on the finding that, the cost of power system components (generator, transformer, submarine cable, and compensator) varies with the frequency and they imbibe a tradeoff in their costs. Specifically, the cost of generator and transformer varies in inverse proportion to frequency. Because, the size of magnetic system (no. of turns/ core area) decreases with increase in frequency and hence cost decreases. Whereas, the cost of submarine cable network varies in direct proportion to frequency. Because, as the frequency increases, the charging current requirement of the cable increases, and for the same active current, the size of the conductor increases with increase in frequency and hence the cost increases.

So it is found that there is tradeoff in two broad categories of cost;

- Sum of cost of generators and transformers
- Sum of cost of submarine cable networks

Figure 2 shows the typical variation of these two costs with respect to frequency. The system cost, which is sum of these two costs, will be parabolic in nature with respect to frequency. The point of inflection on the system cost curve corresponds to the optimum point.

The X and Y co-ordinates of the optimum point corresponds to optimum operational frequency and the optimum cost respectively. The cost of generator and transformer are independent of off-shore distance, but the cost of transmission system depends upon the off-shore distance. So the point of inflection and hence the optimum cost and optimum frequency changes with the distance. Hence it is realized that, for the given power rating and off-shore distance of FFWPS, the system investment cost is a function of frequency and it get optimized (minimized) at certain economically-optimum frequency.

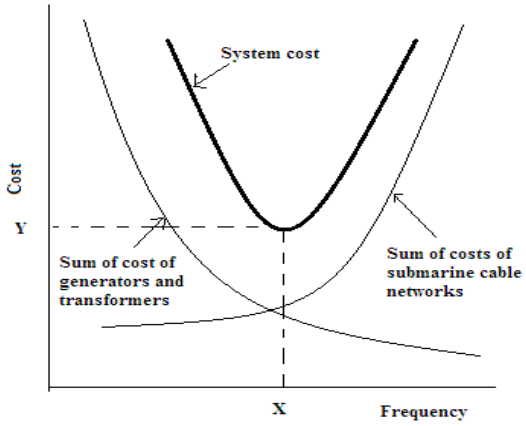


Figure 2 Show the concept of optimum frequency

3. Development of mathematical model

It is evident from the former discussion that, for a given FFWPS, the optimum operational frequency can be determined by optimizing its system cost. It represents an optimization problem. The system cost is sum of the cost of generators, inter-farm transformers, inter-farm electric system and transmission system. Mathematically, it is described as below,

The objective function is defined as,

$$\min\ C(f) \tag{1}$$

where,

$$C(f) = N * C_{gen}(f) + N * C_{itran}(f) + C_{ttran}(f) + C_{in-elec-system}(f) + C_{trans-system}(f) \tag{2}$$

Subject to,

$$1 \leq f \leq 100$$

where,

N- no. of wind turbines in the farm, $C_{gen}(f)$-cost of generator, $C_{itran}(f)$- cost of inter-farm transformer, $C_{ttran}(f)$- cost of transmission transformer, $C_{in\text{-}elec\text{-}system}(f)$- cost of inter-farm electric system, $C_{trans\text{-}system}(f)$- cost of transmission system.

The cost of frequency-converter which is independent of frequency, and the cost of other equipments simply act as an offset, so they are not included.

The cost of generator, inter-farm transformer, transmission transformer, inter-farms electric system and transmission system are estimated as follows.

3.1 Estimating Generator cost

The cost of the generator is estimated from its design parameters. Since the design parameters and hence the cost are

subjective to design constants (variables).To obtain economy, the generator-cost is optimized in terms of its design variables. Specific magnetic loading (B_{av}), specific electric loading (ac), width of ventilating duct (w_{ds}), width of stack length (w_{ss}), slot pitch (y_s), current density (δ) and Short Circuit Ratio (SCR) are selected as design variable with appropriate bonds. Copper and iron are major elements in the generator. So the cost of generator is estimated by considering only the cost of copper and iron. Mathematically it is described as below,

The objective function is defined as,

$$\min \quad C_{gen}(B_{av}, ac, w_{ds}, w_{ss}, y_s, \delta, SCR) \tag{3}$$

where,

$$C_{gen} = C_{ig} + C_{cg} \tag{4}$$

Subject to,

$$0.52 \leq B_{av} \leq 0.65$$
$$20000 \leq ac \leq 40000$$
$$8 \leq w_{ds} \leq 10$$
$$40 \leq w_{ss} \leq 80$$
$$10 \leq y_s \leq 50$$
$$3 \leq \delta \leq 5$$
$$1 \leq SCR \leq 1.5$$

where,

C_{gen} – cost of generator, C_{ig} – iron cost, C_{cg} – copper cost.

Cost of iron and copper is estimated from the volume of iron and copper, which are calculated from the main dimensions, number and size of stator slots, depth of stator core, dimension of rotor pole, depth of rotor core and number of stator and rotor turns.

3.2 Estimating Inter-farm transformer cost

The cost of the transformer is estimated from its design parameters. Since the design parameters and hence the cost are subjective to design constants (variables). To obtain economy, the transformer-cost is optimized in terms of its design variables. A constant (K), as defined in Equation (6) is selected as design variable.

$$k = 4.44 * f * \left(\frac{\varphi}{AT}\right) * 10^3 \tag{6}$$

Cost of copper and iron are only considered for estimating the cost of transformer. Mathematically it is defined as below,

The objective function is defined as,

$$\min \quad C_{itran}(k) \tag{7}$$

where,

$$C_{itran} = C_{it} + C_{ct} \tag{8}$$

Subject to,

$$0.1 \leq k \leq 1$$

where,

C_{itran}– transformer cost (M €), C_{it} - iron cost (M €), C_{ct} - copper cost (M €).

Cost of iron and copper is estimated from the volume of iron and copper, which in-turn are calculated from the core area, primary and secondary turns and width and height of core.

3.3 Estimating Transmission transformer cost

The cost of transmission transformer is estimated as similar to inter-farm transmission transformer (discussed above).

3.4 Estimating Inter-farm electric system cost

The inter-farm electric system includes the interconnection of wind turbines in the farm by submarine cables. The cost of inter-farm electric system is estimated as per by Equation (9). It includes only the cost of copper core. As per the standard practice, size of the core is assumed 50% more than the actual requirement.

$$C_{trans-system} = 3 * \frac{1.5 * I}{\delta} * L * S_c * D_c \tag{9}$$

where,

I – net current to be carried by the cable/phase (amps) , δ – current density (amp/m^2), L – length of the cable (m), S_c- specific cost of copper (/kg), D_c –density of copper (kg/m^3).

The net current is calculated by adding the active current and the charging current as per Equations (10)-(12).

$$I = \sqrt{({I_a}^2 + {I_c}^2)} \tag{10}$$

$$I_a = \frac{P}{(\sqrt{3} * V * \cos \emptyset)} \tag{11}$$

$$I_c = \left(\frac{V}{\sqrt{3}}\right) * 2 * \pi * f * C \tag{12}$$

where,

I_a – active current /phase(amps), I_c – charging current/phase(amps), P – 3-phase power (W), V – line-to line voltage (V), $cos\varphi$ – power factor of load, f – frequency (Hz), C – Capacitance of the cable (F).

3.5 Estimating Transmission system cost

The transmission system exports power from local off-shore grid to on-shore grid by submarine cable. Its cost is estimated as similar to inter-farm electric system (discussed above).

4. Implementation of mathematical model

The mathematical model developed in the previous section, describes the requirement of optimization of system cost model to determine optimum operational frequency. Further, to achieve overall economy, it demands for optimal design of generator and transformers. To implement this, a computational structure with two-stage optimization is developed as shown in the Figure 3. In the first-stage (stage-I), the cost of generator, inter-farm transformer and the transmission transformer are optimized in terms of their design variables, whereas, the system cost is optimized in terms of frequency in the second-stage (stage-II).

This unique feature ensures over-all optimization of the problem. Genetic Algorithm (GA) is used for all the optimizations.

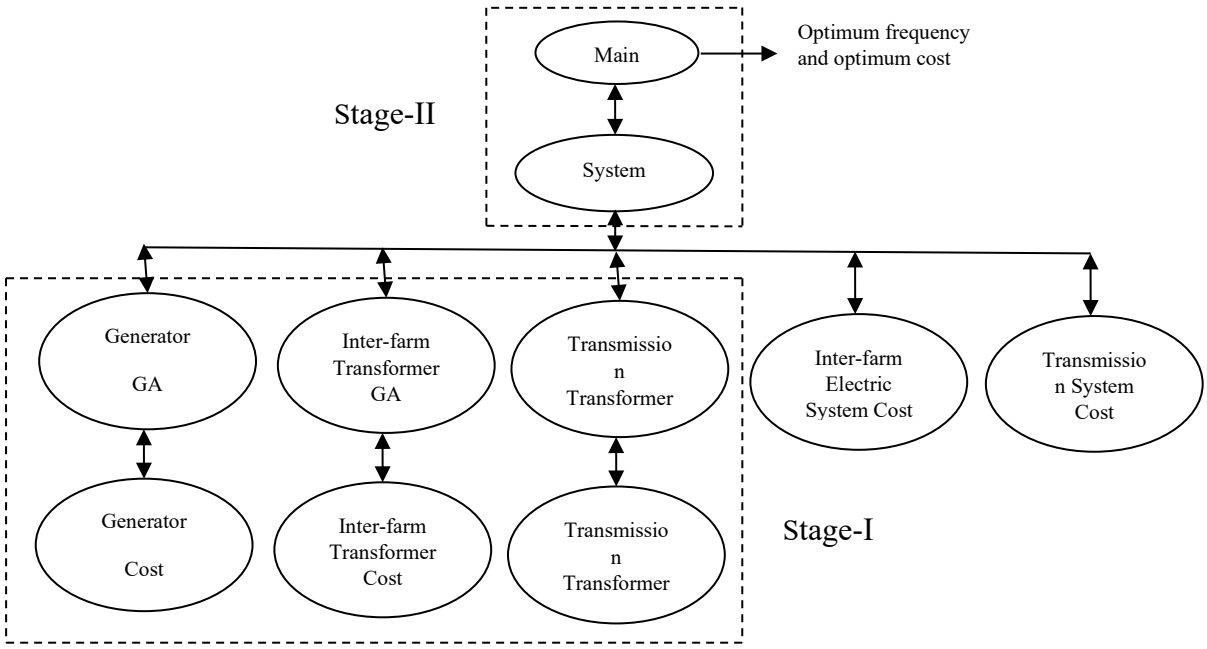


Figure 3: Computational structure of optimization

Figure 4 depicts the process of computation of optimum operational frequency. For simplification, working of Main-GA is only presented. The size of population (fre_pop_max)) for variables (frequency and voltage) and maximum generation limit (gen_max) are set to 50 and 100 respectively. The main core of the algorithm is the evaluation of fitness values (system costs). Initially the population for frequency and voltage is generated randomly. Then, the fitness values are evaluated for each of the population by adding the cost of power system components, which are computed from the executions of respective subroutines. The population is improvised through reproduction, crossover and mutation process and the iteration of generation is continued till the maximum generation count is reached.

Then finally, the minimum fitness value of the last generation which corresponds to the optimum cost is evaluated. The population corresponding to the optimum cost is the optimum frequency.

This methodology of computation of optimum frequency through single-execution effectively addresses all the issues raised by the reviewers in the original paper.

Table 1: Rating of various components of OWPP

Component	**Rating**
Generator	EESG: 2MVA, 690V, 35rpm
Inter-farm transformer	3-phase, core type: 2MVA, 0.69/33kV.
Inter-farm electric system	Submarine cable: 33kV,100km
Transmission Transformer	3-phase core type: 160MVA, 33/150kV
Transmission system	Submarine cable: 160MW,150kV, varying length

5. Results and discussions

Based on the computational structure and the flow chart presented in the previous section, code is written in MATLAB-7.The ratings of power system components are initialized as per the prototype FFWPS. The code is executed under different scenarios to obtain the following results.

6.1 Validation of the generator and transformer design

Electrically Excited Synchronous Generator (EESG) of 2MVA, 690V, 35 rpm is designed using Genetic Algorithm for different frequencies. Table 2 in Appendix, shows the design and performance parameters. Similarly, Table 3 in Appendix shows the design and performance parameters of 3-phase, core type, 2MVA, 0.69/33kV transformer designed using Genetic Algorithm. It can be seen that, core and copper volume for generator and transformer increases with decrease in frequency and hence the cost increases, it complies with the fundamental fact that, the size of the magnetic system increases with decrease in frequency. Since only the cost of iron and copper are taken into account in estimating the cost, the figures given in the tables do not reflect the real costs. Further, since the feasibility of frequency-based design of electric system is evaluated in comparison to LFAC and HVAC, only qualitative variation in the cost with respect to frequency is needed.

6.2 Verifying the existence of optimum frequency

It was postulated in the Section 2 that, the system cost is parabolic in nature w.r.t. frequency and the co-ordinates of the optimum point corresponds to the optimum cost and the optimum operational frequency. To investigate this phenomenon, prototype FFWPS with 50km off-shore distance is considered. The cost of power system components and the system cost are estimated for different frequencies. Table 4 gives the variation in the costs w.r.t. frequency. It can be seen that, the cost of generator, inter-farm transformer and transmission transformer decreases with increase

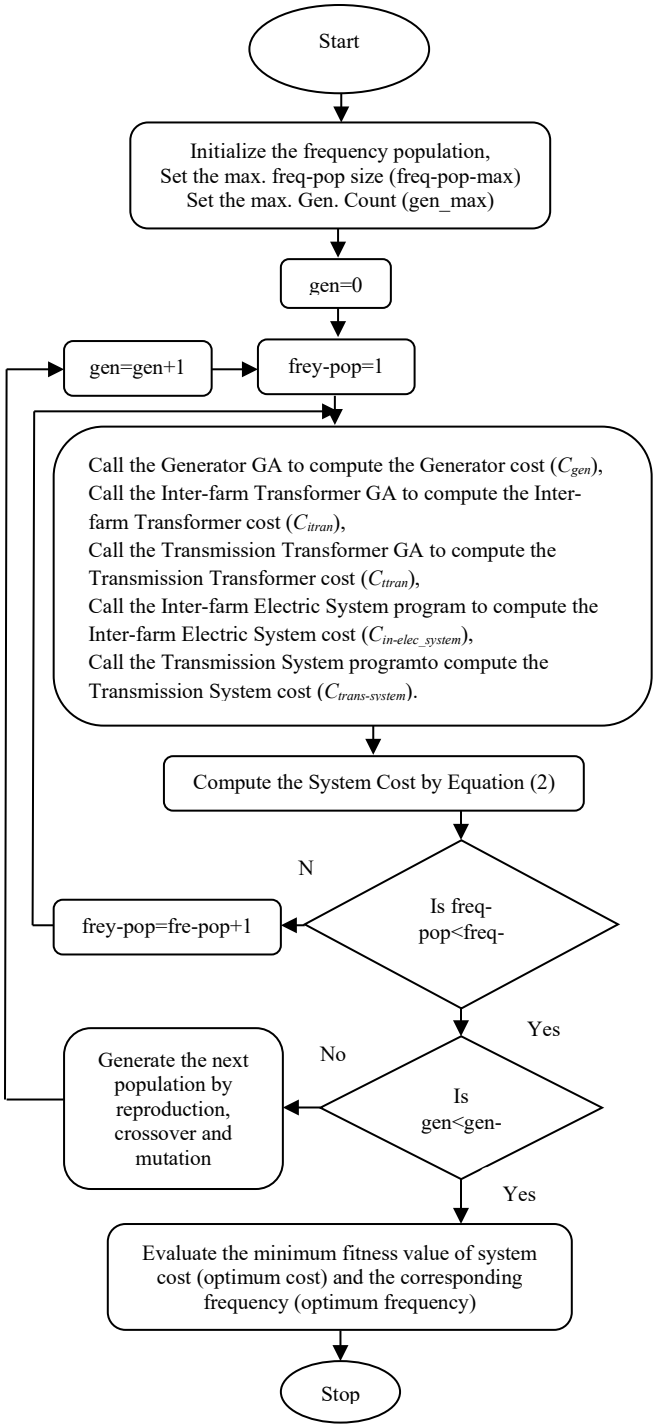


Figure 4: Flow chart showing working of computational structure

in frequency, whereas the cost of inter-farm electric system and transmission system increases with increase in frequency. It can also be seen that, the system cost decreases initially up-to certain optimum frequency (indicated by shading) and then increases with the frequency. Figure 5 shows the system cost w.r.t. frequency. It can be seen that, the cost of FFWPS is parabolic in nature and their lies an optimum operational frequency where the system cost is minimum. For 50km transmission length, the optimum operational frequency is obtained as 52Hz and the corresponding optimum cost is 64.73M€. If the system is operated at 50/3Hz, the system cost will be 78.43M€. Thus the result clearly indicates that, if the FFWPS is operated at optimum operational frequency rather than operating at 50/3Hz, substantial amount of saving can be achieved. Figure 6 shows the savings for different frequencies. It can be seen that; the saving is maximum (17.47%) at the optimum operational frequency. Thus it is verified that, for the given FFWPS, there exist an optimal operational frequency.

Table 4: Variation in the cost of various power system components w.r.t. frequency

(1)	(2)	(3)	(4)	(5)	(6)	(7)
50/3	21.31	8.12	19	0.78	29.21	78.43
20	18.25	6.75	19.14	0.68	29.31	74.14
30	13.39	4.58	19.7	0.52	29.7	67.89
40	10.75	3.5	20.45	0.42	30.25	65.37
52	9.22	2.84	21.38	0.35	30.94	64.73
60	8.2	2.43	22.46	0.31	31.77	65.18
70	7.56	2.11	23.68	0.28	32.72	66.34
80	6.99	1.86	25.01	0.26	33.78	67.9
90	6.51	1.68	26.43	0.24	34.95	69.8
100	6.05	1.59	27.94	0.22	36.21	72.01

(1)- Frequency (Hz), (2)- Generator cost (M€), (3)- Inter-farm transformer cost (M€), (4)- Inter-farm electric system cost (M€), Transmission transformer cost (M€), (5)- Transmission system cost (M€), (7)- System cost (M€)

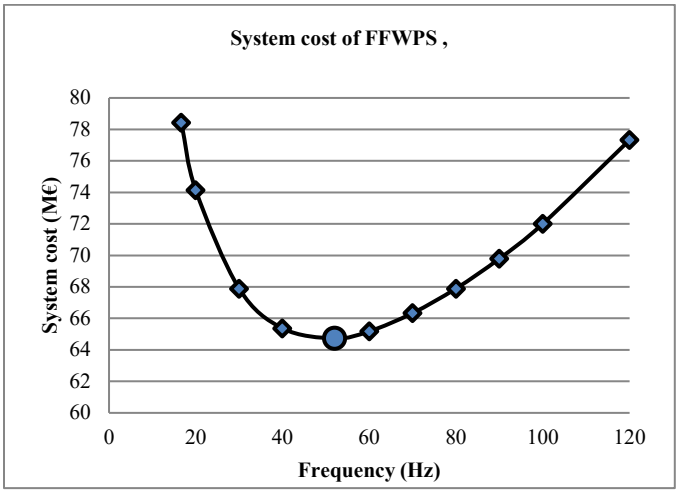


Figure 5: Variation in cost of FFWPS v/s frequency (off-shore distance 50km)

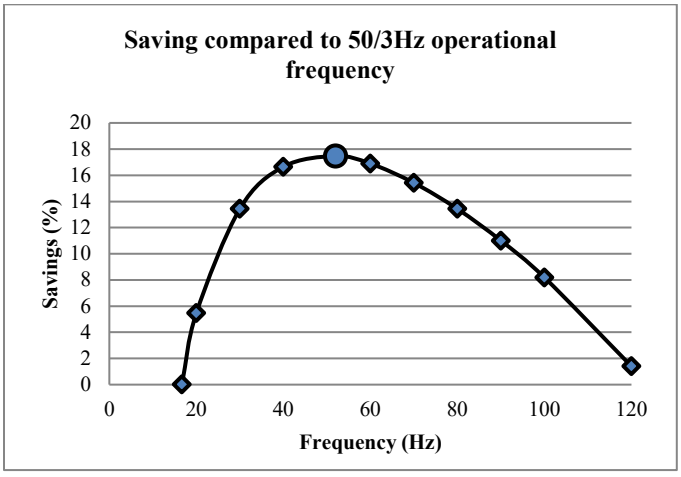


Figure 6: Saving with optimum design in comparison to LFAC (off-shore distance 50km)

Further, since the optimum operational frequency depends upon the off-shore distance. With the aid of **Main-GA,** optimum operational frequencies for different transmission lengths are obtained as shown in the Figure 7. It can be seen that, the optimum operational frequency decreases with increase in the transmission length. It complies with the fact that, HVAC (50 Hz) is economical for short distance transmission and HVDC (0 Hz) is for long distance.

5.1. Feasibility of optimum frequency-based design

As stated earlier, the main objective of paper is to enhance the project economy by reducing its electric infrastructural cost through its operation at optimum frequency. To investigate this, savings accrued from optimum frequency-based design, in comparison to conventional HVAC (50Hz) and LFAC (50/3Hz) are estimated as shown in the Figure 8 for different transmission lengths.

The saving is expressed as percentage of the cost of system under comparison. It can be seen that, compared to LFAC, optimum frequency-based system gives substantial amount of saving in the investment cost up-to 150km (approx.). Whereas, compared to HVAC, the saving is substantial beyond 80km (approx.). So, when it is question of choice of system, the results indicates that, HVAC is undoubtedly better up-to 80km distance, but beyond that and up-to 150km, optimum frequency-based system is a promising choice. The results suggest that, for the given FFWPS, optimum frequency-based system is feasible for the off-shore distance in the range of 80-150 km, and for this range the operational frequency is 20 to 40 Hz.

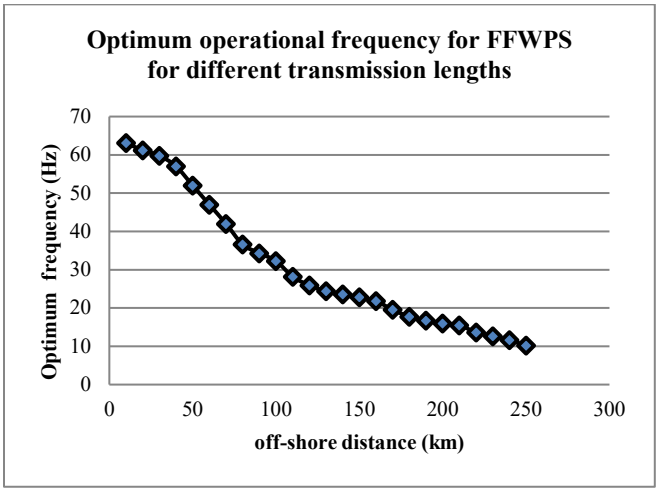


Figure 7: Optimum operational frequencies for FFWPS for different off-shore distances

6.2 Consideration of optimum operating voltage

The savings shown in the Figure 8 are obtained with fixed operating voltages i.e. 33kV for inter-farm electric system and 150kV for transmission system. But the savings can be enhanced enormously if the operating voltages are selected appropriately. The basic thumb rule recommends for the selection of highest possible operating voltages. Even-though, higher operating voltage reduces the cost of inter-farm electric system and transmission system, but on contrary, they increase the cost of inter-farm transformer and transmission transformer.

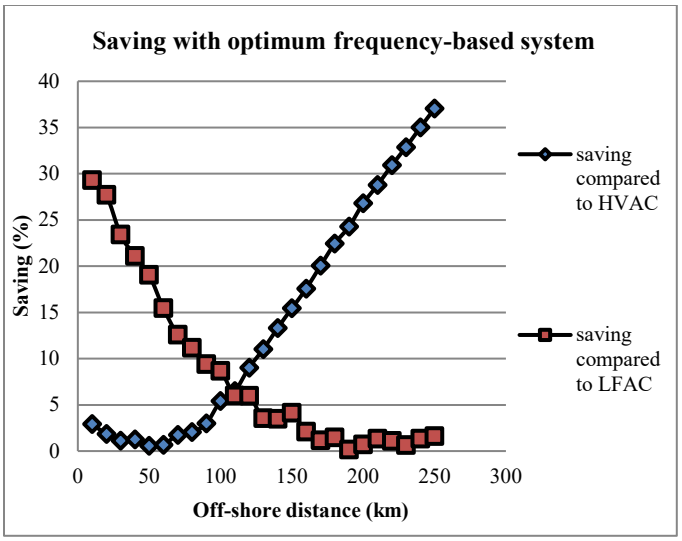


Figure 8: Saving with optimum frequency-based system

So there is trade-off in the costs and it demands for optimum selection of operating voltages.

To address this issue, the operating voltages of inter-farm electric system and transmission system are initialized as variables in the optimization (Main-GA) along with frequency. This leads to an idea of optimum frequency-voltage-based design of FFWPS. Figure 9 shows the saving accrued from this system compared to HVAC and LFAC. For HVAC and LFAC system the operating voltages are kept fixed as before.

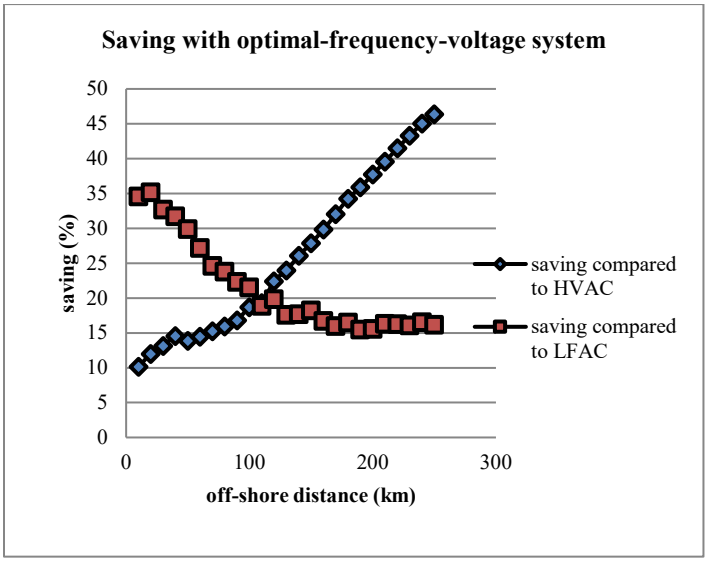


Figure 9: Saving with optimum-frequency-voltage system

It can be seen that the savings are augmented by optimum selection of operating voltages along with optimum choice of operational frequency.

Thus the results reveal that, there exits an optimum operational frequency and if the electric system is designed and operated at the optimum frequency rather at the standard conventional frequencies, substantial amount of saving in the investment can be achieved.

6. Conclusions

The paper has extended the work on optimum frequency-based design of electric system for OWPP. It deals with the computation of most economical operational frequency so that the electrical infrastructural cost is minimal. The paper has presented a newer methodology for the optimal design of electric system through optimization of frequency-based cost model of the system in a single execution. The developed methodology is applied to a prototype FFWPS and shown that the investment cost can be reduced significantly if the system is designed and operated at the optimum frequency. Further, the paper effectively addresses all the issues of the original paper. It is concluded that, the proposed optimal design for electric system is economically competitive compared to HVAC and LFAC for moderate transmission distance. The work carried out in this paper provides a new dimension for economical design of electric system. It also helps in motivating the investors for deployment of OWF to dovetail the energy gap between generation and demand. Further it also indirectly addresses the problem of environment pollution and global warming which are the major concerns of world today in the endeavor. Nevertheless in the proposed design, power system components such as generator, transformer and compensation units need to be redesigned for the operation at nonconventional frequencies. But for large size wind farms (>1000MW), it will not be major paramount because the required generators and transformers can be manufactured cheaply in mass. In the present paper only investment cost is considered for the optimal design but in the future work operational cost can also be added in the optimization to get better results. Further, the study can be extended to other configurations of electric system.

References

[1] Uttam S. Stpute, D.R. Joshi, "Optimum Frequency-Based Design of Off-shore Wind Power Plant (OWPP)", proc. of 2017 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Bengaluru, India, Nov. 2017.

[2] 'Global Wind Energy Council (GWEC) global 2015 report', available at : http://www.gwec.net/global, visited on: 25th August 2018.

[3] Yuan-Kang Wu, Li Wang, Yong-Qing Huang, Shu-Wei Liu, "Overview of Important State-of-the-art Technologies in Off-shore Wind Energy System," International Journal of Smart Grid and Clean Energy, vol. 2, no. 2, pp. 64-67, May 2013.

[4] X. Wang, "The fractional frequency transmission system," in Proc. IEEE 5th Annu. Conf. Jpn. Power Energy, Tokyo, Japan, Jul. 1994,pp 53-58.

[5] X. Wang and X. Wang, "Feasibility study of fractional frequency transmission system," IEEE Trans. Power Syst., vol. 11, no. 2, pp. 962-967, May 1996.

[6] X. Wang, C. Cao, and Z. Zhou, "Experiment on fractional frequency transmission system," IEEE Trans. Power Syst., vol. 21, no. 1, pp. 372-377, Feb. 2006.

[7] U. Satpute, S. Jangamshetty, and D. Joshi, "Feasibility study of fractional frequency transmission system," in Proc. Joint Int. Conf. Power Electron. Drives Energy Syst., New Delhi, India, Dec. 2010, pp. 1-6.

[8] X. Wang, "Fractional frequency transmission system and its applications on wind power and renewable resource integration into system," Power Elect. Eng., vol. 2, no. 1, pp. 12-14, 2008.

[9] W. Fischer, R. Braun and I. Erlich, "Low frequency high voltage offshore grid for transmission of renewable power", proc. of 3rd IEEE PES Innovative SmartGrid Technologies Europe, Berlin, Oct. 2012.

[10] X. Wang, "Fractional frequency wind power system," Mod. Elect. Power, vol. 24, no. 1, p. 1, 2007

[11] Z. Song, X. Wang, Y. Teng, L. Ning, and Y. Meng, "Optimum control study for fractional frequency wind power system," in Proc. Asia-Pac. Power Energy Eng. Conf. Shanghai, China, Mar. 2012, pp. 1-5.

[12] O. Gomis-Bellmunt, A. Junyent-Ferr´e, A. Sumper, and S. Galceran-Arellano, "Maximum generation power evaluation of variable frequency offshore wind farms when connected to a single power converter," Appl.Energy, vol. 87, no. 10, pp. 3103–3109, Oct. 2010.

[13] Xifan Wang, Xiaohui Wei, and YongqingMeng, "Experiment on grid-connection process of wind turbines in fractional frequency wind power system," IEEE Trans. On Energy Conversion, vol. 30, no. 1, pp. 22-31, March 2015.

Appendix

Table 2: Design and performance parameters of 2MVA, 690V, Salient pole Synchronous Generator

Design and performance parameters	**5Hz**	**10Hz**	**50/3Hz**	**20Hz**	**25Hz**	**30Hz**	**35Hz**	**40Hz**	**45Hz**	**50Hz**
Optimal Cost (M €)	0.7355	0.4152	0.2661	0.2291	0.1935	0.1689	0.1490	0.1360	0.1258	0.1166
Stator Bore Diameter (m)	3.9244	4.5426	5.2074	5.5080	5.8600	6.2130	6.4700	6.7840	6.9900	7.2640
Stator Length (m)	0.8219	0.6205	0.4674	0.4220	0.3760	0.3370	0.3080	0.2842	0.2649	0.2481
Poles	18	34	58	70	86	104	120	138	154	172
Turns/pole	51	56	66	69	72	77	80	85	86	88
Flux/pole (wb)	0.3655	0.1668	0.0849	0.0677	0.0519	0.0404	0.0334	0.0275	0.0242	0.0213
Volume of stator iron (m^3)	1.8363	0.8836	0.4656	0.3753	0.2871	0.2267	0.1791	0.1484	0.1259	0.1060
Volume of stator copper (m^3)	0.2026	0.1624	0.1482	0.1426	0.1363	0.1345	0.1320	0.1329	0.1281	0.1258
Volume of rotor iron (m^3)	3.3626	1.8963	1.1727	0.9993	0.8236	0.6951	0.6032	0.5330	0.4825	0.4446
Volume of rotor copper (m^3)	0.2841	0.2210	0.1982	0.1805	0.1720	0.1681	0.1571	0.1527	0.1520	0.1430
Loss (MW)	0.4197	0.2839	0.2289	0.2118	0.1833	0.1739	0.1632	0.1585	0.1539	0.1485
Efficiency (%)	79.400	85.100	87.910	88.840	89.670	90.170	90.690	90.940	91.230	91.540
Stator temp rise (^{0}c)	196.05	141.89	107.00	98.210	88.720	77.610	75.970	68.350	66.194	64.250
Rotor temp rise (^{0}c)	236.32	214.99	190.32	188.73	174.63	164.77	160.75	159.32	144.42	144.54

Table 3: Design and performance parameters of 2MVA, 0.69/33kV, 3-phase core type transformer

Design and performance parameters	**5Hz**	**10Hz**	**50/3Hz**	**20Hz**	**25Hz**	**30Hz**	**35Hz**	**40Hz**	**45Hz**	**50Hz**
Optimal Cost (M €)	0.2350	0.1447	0.0999	0.0838	0.0678	0.0571	0.0496	0.0434	0.0391	0.0354
Constant (k)	0.1150	0.1524	0.3353	0.3556	0.3830	0.4064	0.4350	0.4427	0.4669	0.4801
Core area (m^2)	0.1544	0.1023	0.1351	0.1194	0.1029	0.0910	0.0835	0.0743	0.0697	0.0645
Primary turns	77	60	28	25	25	24	24	21	21	21
Secondary turns	3705	2795	1275	1200	1100	1050	975	950	900	875
Width of transformer (m)	1.9492	1.5867	1.8230	1.7138	1.5907	1.4960	1.4328	1.3521	1.3092	1.2594
Height of transformer (m)	1.3875	1.1294	1.2977	1.2199	1.1323	1.0648	1.0199	0.9624	0.9319	0.8965
Volume of core (m^3)	6970	3760	5700	4730	3790	3150	2770	2330	2110	1880
Volume of copper (m^3)	6180	4230	439	400	359	338	317	309	269	266
Loss (MW)	0.1258	0.0831	0.0717	0.0614	0.0549	0.0513	0.0431	0.0469	0.0407	0.0395
Efficiency (%)	92.7658	95.066	95.858	96.311	96.661	96.895	97.380	97.153	97.519	97.500
Regulation (%)	4.5138	3.4294	19.519	21.298	26.561	32.104	21.080	49.623	38.329	47.159
No-load current as % of full load current (%)	7.2447	5.9768	11.522	12.000	11.134	10.882	12.368	9.8491	10.792	10.386

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 159-164 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Low Contrast Image Enhancement Using Convolutional Neural Network with Simple Reflection Model

Young Shik Moon*, Bok Gyu Han, Hyeon Seok Yang, Ho Gyeong Lee

Hanyang University, Computer Science & Engineering, 15588, Republic of Korea

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 23 January, 2019
Online : 5 February, 2019

Keywords:
Image Enhancement
Convolutional Neural Network
Reflection Model
Machine Learning

ABSTRACT

Low contrast images degrade the performance of image processing system. To solve the issue, plenty of image enhancement methods have been proposed. But the methods work properly on the fixed environment or specific images. The methods dependent on fixed image conditions cannot perform image enhancement properly and perspective of smart device users, algorithms including iterative calculations are inconvenient for users. To avoid these issues, we propose a locally adaptive contrast enhancement method using CNN and simple reflection model. The experimental results show that the proposed method reduces over-enhancement, while recovering the details of the low contrast regions.

1. Introduction

Image enhancement is an important and typical topic of computer vision. Although the performance of digital systems has improved greatly, the low quality of some images due to external factors such as environment and backlight, which may degrade the performance of image processing systems such as intelligent traffic systems, visual surveillance, and consumer electronics [1]. Especially, low contrast images reduce visibility. Therefore, many smart devices help users take pictures and improve their results with internal image processing methods, but require some input, such as a low contrast area input by the device user, or an average brightness and hue. In specific condition, it is a good solution for the device user who does not consider about the visibility of the image such as counter light images on purpose. However, only chooses the color tone and increases overall brightness is not fit for image enhancement in view of the simple device users.

While maintaining the image quality, to improve the low contrast region of input image is not easy. To solve the problem in this conditional situation, a plenty of methods have been proposed based on mathematical knowledge and traditional image processing method but the methods only perform well in fixed circumstance or make side effect such as over-enhancement or halo effect. To reduce the undesired effect, some enhancement methods use optimization techniques with image decomposition mechanism. However, the computation time is increased and the methods may produce unintended adverse effects depending on the objective function with decomposed factors. Because of these disadvantage, the methods are not proper to simple device users.

In the past, artificial neural network methods have been one of the most difficult methods to obtain successful results due to small amounts of computation resource and data. However, artificial neural networks are widely used in many fields due to the development of devices and explosively increased amount of data. In addition, as many learning methods are studied, many effective approaches based on input data conditions have been proposed. Recently, the deep learning based methods show the best performance in computer vision sections. The methods are widely being employed on many computer vision tasks that are not easily solved with traditional methods such as image classification, image segmentation, etc. The deep learning based methods with well-labeled training data automatically abstract features of the image differently from the traditional methods and reconstruct the final result with the extracted features. However, the performance of the methods is built on well-labelled data. To overcome the conditional limits, several training techniques are proposed such as semi-supervised learning and weakly-supervised learning. From the perspective of performance, using these skills are able to lead to better results and it can be used widely in general purpose. The performance of the deep learning models trained using non-labelled or rough labelled data shows the power of representation ability.

Focus on the advantage of deep learning, we propose a locally adaptive contrast enhancement method. The method follows previous image enhancement mechanism basically, detecting low contrast region and apply enhancement algorithm. Each person

*Young Shik Moon, 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do, Republic of Korea, ysmoon@hanyang.ac.kr

evaluations for the low contrast region are different. Therefore, we use deep learning mechanism to employ the advantage which extracting feature factor from dataset automatically. Even though our method is based on training with confusing data, the power of representation ability of deep learning shows well results on contrast enhancement purpose. Also, we use a simple reflection model to reduce computational burden.

The remainder of this paper is structured as follows: Section 2 provides background information on image enhancement and briefly introduces previous methods. Section 3 gives a detail of proposed method and shows the detail of our network. While Section 4 provides the experiment results of image contrast enhancement with the proposed method and the previous methods. Finally, Section 5 concludes this paper.

2. Related works

For image enhancement, a lot of image processing methods were proposed previously. Histogram based methods are the most popular due to its simplicity and effectiveness. Basically, histogram equalization (HE) is a well-known traditional contrast enhancement method. The main idea of HE contains automatic calculation of the uniform histogram distribution in dynamic range. But HE method does not preserve to mean intensity of input images. The mean intensity of an input image and result of HE are different and it cause undesired effect such an over-enhancement. To solve the issue, brightness preserving bi-histogram equalization (BBHE) [2] was proposed. It divides histogram of an input image into sub histograms and enhances contrast separately. Adaptive histogram equalization (AHE) is used for enhancing contrast in images. It differs from HE by adaptive method that computes several histograms and each histogram corresponding to a distinct section of an image. AHE divide an input image into small sub images and apply HE to each sub image. But because of the size of the sub image, it is too sensitive to noise. Contrast-limited adaptive histogram equalization (CLAHE) [3] is improved version of AHE. The processes of CLAHE is basically similar to AHE. CLAHE controls the sensitive noise with contrast limited distribution parameters. The neighboring tiles, sub image, are combined by bilinear interpolation. Especially in homogeneous areas, the contrast can be limited to avoid noises by distribution parameter.

Retinex-based methods assume that recognizes the relative brightness of the scene, rather than recognizing the brightness of the scene at a certain position when the human visual system recognizes the scene. Retinex based methods have been proposed widely and single-scale retinex (SSR) [4], multi-scale retinex (MSR) [5] and multi-scale retinex with color restoration (MSRCR) [6] are most well-known algorithms. SSR is a method of using the difference between the center pixel value and the convolution result around the center pixel on the log scale. Several convolution masks were proposed [7, 8] but to select the variance of convolution mask effects largely and inadequate variance value can be lead to deterioration of image quality. The MSR is proposed as a method to mitigate this problem of SSR and the retinex result is calculated as the weighted average value of SSR with multiple variance values. The MSRCR proposed that color restoration function (CRF) added to MSR result to solve a problem that come out the gray scale result if specific color is dominant in the input image.

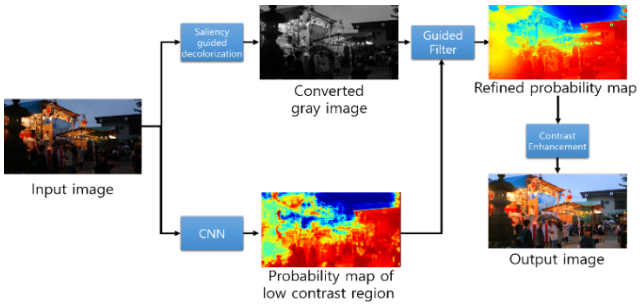


Figure 1. Overall flow of the proposed method.

In recent years, some methods are proposed to enhance low contrast images with adjustment to decomposed illumination image which obtained from various feature models. Different feature models lead to different illumination images and affect the enhanced images. R. Chouhan [9] proposed image enhancement method based on image decomposition technique which follow the discrete wavelet transform (DWT). X. Fu [10] proposed a probabilistic method based on image decomposition. The method decomposed input image into illumination and reflection model then applied the maximum a posteriori (MAP) for estimating enhanced illumination and reflection effectively. Y. ZhenQuiang [11] proposed low-light image enhancement using the camera response model. The camera response model consist of two sub models, camera response function (CRF) and brightness transform function (BTF). Based on these models and estimated exposure ratio map, they enhance image each pixel of the low light image.

The convolutional neural network (CNN) [12] proposed by Y. Lecun has been used widely in computer vision. CNN is being employed on many hard tasks and it showed the best results in many competition in the area. Basically, neural network was proposed for estimating test samples based on training data with hidden layers but calculation with just hidden layer is not suitable for some tasks which related to 2D and 3D data such as images and videos. To preserve locality and flexibility, the Y. Lecun proposed

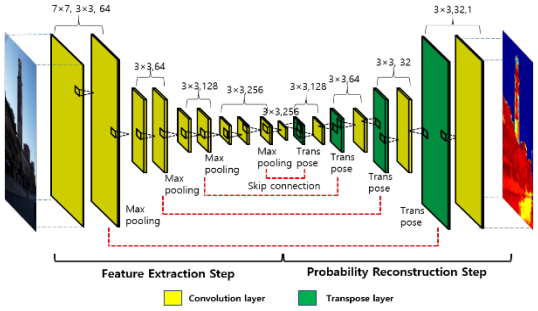


Figure 2. Structure of the convolutional neural network to detect low contrast regions.

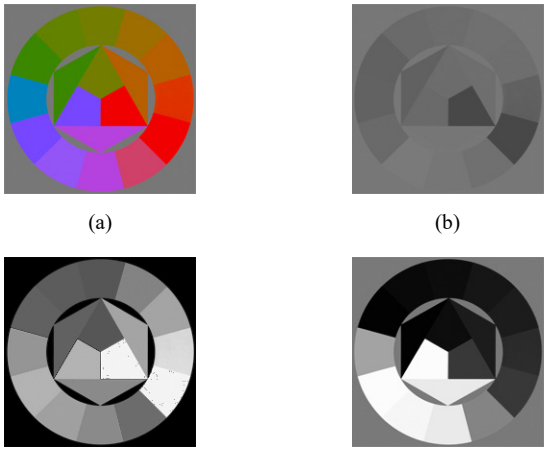
(a) (b)

Figure 3. Example of training data.
(a) Original image. (b) Labelled image.
basic mechanism of CNN with simple convolutional operation.

One of the most improved tasks based on deep learning in computer vision task is semantic segmentation. Traditional segmentation algorithm was based on mathematical knowledge such as specific feature extraction or the relationship of each pixel. Because of the these factors it required in deep research each object which to be segmented. The features that can segment each of object in image are used in restrictive environment. But the methods based on deep learning, It works well without the need to handcraft features of the image based on network characteristics and training data. The most well-known model in semantic segmentation based on deep learning is fully convolutional networks (FCN) proposed by E. Shelhamer [13]. This method shows a basic segmentation network architecture with convolution, pooling and deconvolution. In an input image, important feature information is extracted and compressed by the trained mask through convolution layer and the pooling layer. The compressed information is upsampled in final step with deconvolution layer for matching the size of the input image and the result image. This network not only contains a basic idea for segmentation but also skip-connection mechanism through experimental results. H. Noh [14] proposed a segmentation method with deconvolution technique. The basic model is encoder-decoder network and each lost spatial data by pooling is compensated by deconvolution and unpooling layer.

3. Proposed method

The proposed method consists of four steps in total. The first step is low contrast estimation step that generate low contrast probability map. The second step is a create contrast gray scale step. The third step is a refining probability map step with obtained probability map and converted gray scale image. The final step is an enhancement step that enhance low contrast image with refined probability map and converted gray scale. Figure 1 shows the flow chart of the proposed method.

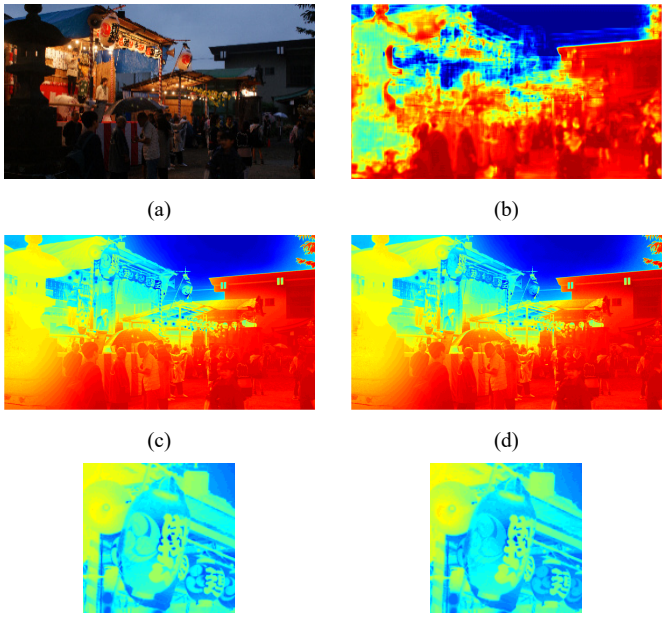


(a) (b) (c) (d)

Figure 4. Comparison of gray scale images.
(a) Input image. (b) Standard gray image.
(c) Contrast chromatic weight. (d) Contrast gray image.

3.1. Low contrast region probability map

The key idea of the proposed method is to generate probability map of low contrast region. For generation of the probability map quickly and working on small devices such as mobile and tablet PC without computation burden, we construct the network as small as possible. The Figure 2 shows structure of the proposed method. As shown Figure 2 the network consist of 8 convolution blocks. Inspired by semantic segmentation networks, we design the structure similar to these networks, encoder-decoder networks.



(a) (b) (c) (d) (e) (f)

Figure 5. Comparison of refined probability maps.
(a) Input image. (b) Probability map.
(c) Refined the (b) with standard gray image.
(d) Refined the (b) with contrast gray image.
(e) Part of (c). (f) Part of (d).

From first convolution block to transpose layer, it is called feature extraction step which abstracts low contrast region's feature. All the remaining layers belong to image reconstruction step which rebuild probability map of low contrast regions. To compensate for the lost spatial information by using pooling layer, we use skip connection mechanism and transpose layers. The skip connection was proposed to improve performance of semantic segmentation network that it added the extracted spatial information from the extraction step to the reconstruction step.

We have train the network to generate probability map of low contrast region with low contrast image dataset which collected from on the internet and we used published dataset [15]. We have made ground truth manually and each label consisted of 3 label colors(red, yellow, none). The colors is converted to 1, 0.5 and 0

(a) (b)

Figure 6. Result of enhancement step.
(a) Input image. (b) Final result of enhancement step.

for training step. To get enough training data, we have applied data augmentation including rotation (-10, 0, and +10 degree) and mirroring (vertical, horizontal, and diagonal). As result, the size of training data is 6924. Figure 3 shows example of training data.

3.2. Contrast gray image

In the previous works, [1, 16] used Chromatic contrast weights [17] and standard gray scale image for refining the probability map. But only use standard gray scale was not proper because of the image contrast. The standard gray scale image is built with constant ratio of each channel therefore the image was not be able to contain the contrast of each object's shape. To solve the issue, using the chromatic contrast weight was proposed with handled constant values such as angle of HSV offset and color opponent. The chromatic contrast weight was calculated based on these values, they showed similar values even difference color. To convert an input color image to the gray image appropriately, we use saliency-guided decolorization method [18]. The method based on sparse model and it is able to express contrast with gray value properly. Figure 4 shows gray convert result of standard, chromatic weight and saliency-guided decolorization method.

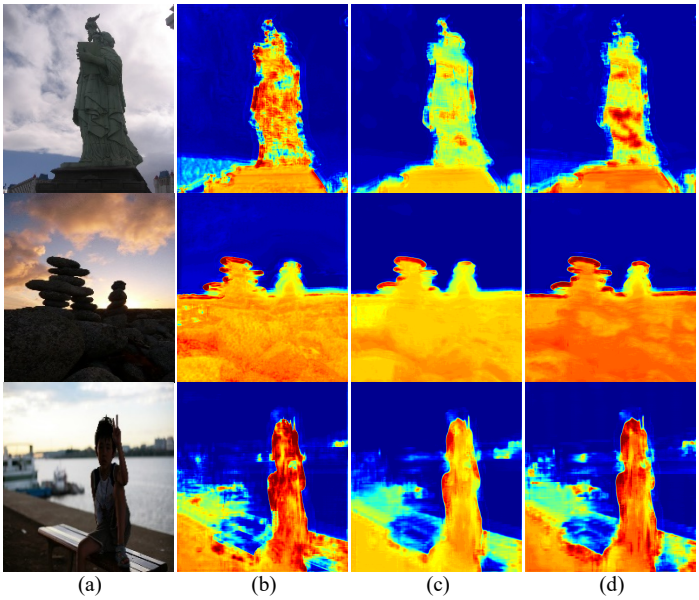
3.3. Refining outline of probability map

In this step, we refine the probability map with guided filter [19] and converted contrast gray image. The guided filter performs as edge-preserving smooth filter which is well known filter in the computer vision area. The guided filter needs the guided image and filtering image, we chose the probability map as a filtering image and the contrast gray image as guided image. It has performed that the boundary region of the probability map refined similar to contrast gray image. Refined the boundary region and shape prevent halo effect and undesired over-enhancement. Figure 5 shows the result of this step. In (e) and (f) of Figure 5, the refined probability map based on standard gray image shows lost detail shape but (f) shows the detail of tomoe pattern.

3.4. Enhancement

The final step is enhancement step. In this step we follow the reflection model :

$$I(x) = L(x)J(x)+w \quad (1)$$



(a) (b) (c) (d)

Figure 7. Comparison of the probability maps.
(a) Input image. (b) Probability map of proposed method.
(c) Without skip connection. (d) No pooling.

where $I(x)$ is an observed image, $J(x)$ is a desired(enhanced) image, $L(x)$ is illuminance map which is a degraded operator, and w is additive noise. We assume that the image has no noise. So, we can get the enhanced image following (2)

$$J(x) = \frac{I(x)}{F(L(x))+\epsilon} \quad (2)$$

where F is the guided filter and ϵ is small value which is used to avoid dividing by zero. For implementation of (2), we used (2) as

$$\text{Output}^{ch}_{enhanced} = \frac{I^{ch}(x,y)}{(1-Refined\ map(x,y))+\epsilon} \quad (3)$$

where I^{ch} is each color channel of the input image, and the refined map is the result computed in the previous step. Figure 6 shows result of the enhancement step.

4. Experiment and results

In the following section we evaluate performance of the proposed method. In Section 4.1, we compare the network performance of the proposed method by comparing two different networks. The performance of the proposed method is then evaluated by comparing the results of the proposed method and the previous methods. The experiments were performed on a PC with a 3.20 GHz Intel Pentium Quad Core Processor and GTX 1080ti graphics card.

4.1. Network structure experimentation

For the comparison of effect of the networks, we made two kind of networks. First network was built without skip-connection and the second network was constructed without pooling layers. Each networks were trained same setting as proposed method such as hyper parameters, size of kernel, and number of channels. The Figure 7 shows the generated probability map of each network. The probability maps in Figure 7, the proposed method can get better shape than the others because of reconstruction power in decoder section. Also, the results of no-skip network and no-pooling networks can be seen quite similar. However, from a detailed point of view, the result of no pooling network shows better shape than no-skip network because of the lost the spatial information by pooling layers. In the previous work [13], they proved the power of skip connection and pooling layer but they used precise labeled data for solving each task. Basically, the neural network is used for estimation of some test data based on training data. Even though the training data we used were not precise, we reconfirmed that the network architecture and the inner method could drive quite successful result.

4.2. Comparison with other methods

In order to evaluate the performance of the proposed method, we compared the performance of the proposed method and the previous methods, CLAHE [3], MSRCR [6], Fu's method [10], and Ying's method [11]. Figure 8 shows the results of each methods. The first row of Figure 8, the test image was taken from outside which contains the light source and low contrast region such as the counter light region. In the result of CLAHE, the contrast was all most same as input image. Because of the input image contains light source and contrast region together, the size of sub image is sensitive and the distribution of the sub image is controlled by the distribution parameter. The result by MSRCR

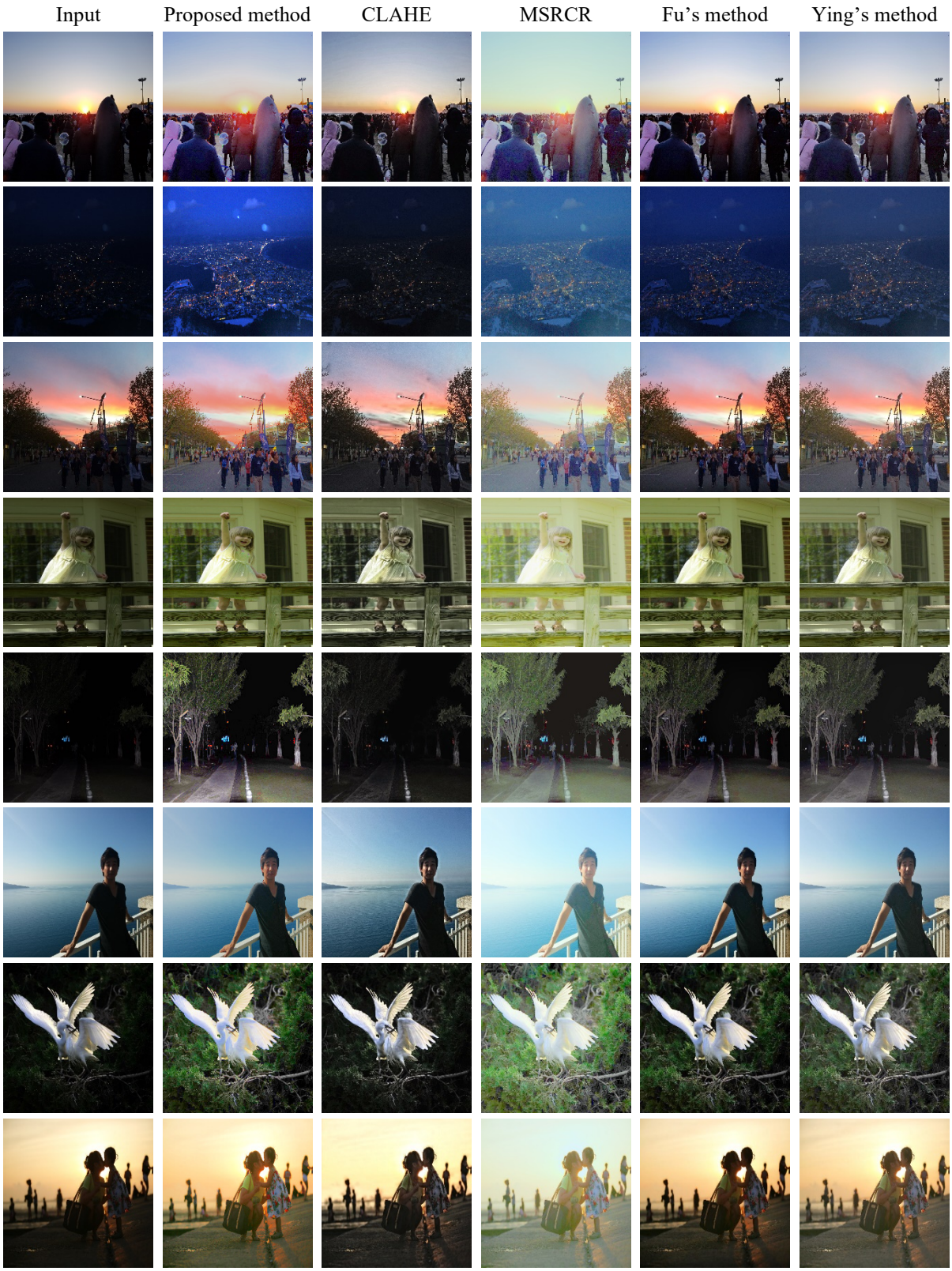


Figure 8. Results of proposed method and previous methods.

(a) (b) (c)

Figure 9. Zoom in the result of Fu's method, Ying's method and proposed method.
(a) Proposed method. (b) Fu's method.
(c) Ying's method.

that not only the contrast is better than CLAHE but also the shape of the statue is seen clearly. But overall brightness is increased too much and the color of the sky looks the same color. The results of Fu's method and Ying's method are improved properly in perspective of contrast and shape. But the overall brightness results are inadequate. In the second row of Figure 8, the results of overall brightness are quite dark. Therefore, the results cannot show the detail of the sky view. For more accurate qualitative comparison, Figure 9 shows zoomed in the results of Fu's method, Ying's method, and proposed method. As you can see images in Figure 9, the shade of the man's face is enhanced properly and we can distinguish the boundary between the nose and the cheek.

5. Conclusion

In this paper, we proposed the image enhancement algorithm with simple reflection model and CNN. In order to enhance low contrast image, plenty of methods have been proposed and result of these methods show good quality from perspective of researcher. But previous methods showed undesired effects such as over- enhancement or unnatural results in unfit circumstance such as illumination setting and image sizes. To solve these problems, CNN was employed to detect low contrast regions in order to preserve the shape of low contrast region and prevent over-enhancement. For the purpose of expression the shape of input images, we employed the saliency-guided decolorization method and the guided filter. We showed our network was suitable for image enhancement through experimental network changes and results, and proved that the results of the proposed method were superior to those of the previous methods through comparison of results with the methods. As a result, the proposed method showed successful enhancement without side effects.

References

[1] B. K. Han, H. S. Yang, and Y. S. Moon, "Locally Adaptive Contrast Enhancement Using Convolutional Neural Network," IEEE International Conference on Consumer Electronics, Las Vegas, USA, 2018. http://doi.org/10.1109/ICCE.2018.8326096

[2] Y. T. Kim, "Contrast Enhancement Using Brightness Preserving Bi-Histogram Equalization," IEEE Trans., 1997. http://doi.org/10.1109/30.580378

[3] K. Zuiderveld, Graphics Gems IV, Academic Press, 1994.

[4] D. Marini and A. Rizzi, "Computational approach to color adaptation effects," Image and Vision Computing, 18(13), pp. 1005–1014. Oct. 2000. https://doi.org/10.1016/S0262-8856(00)00037-8

[5] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A Multi-Scale Retinex For Bridging the Gap Between Color Images and the Human Observation of Scenes," IEEE Trans. on Image Processing: Special Issue on Color Processing, 6, pp. 965-976, 1997. http://doi.org/10.1109/83.597272

[6] Z. Rahman, G.A. Woodell, and D.J. Jobson, "Multiscale Retinex for Color Image Enhancement," Proceedings of 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 1996.

[7] E. Provenzi, M. Fierro, A. Rizzi, L. De Carli, D. Gadia, and D. Marini, "Random spray retinex: A new retinex implementation to investigate the local properties of the model," IEEE Transactions on Image Processing, 16(1), pp. 162–171, Jan. 2007. https://doi.org/10.1109/TIP.2006.884946

[8] D.J. Jobson, Z. Rahman, and G.A. Woodell, "Properties and performance of a center/surround retinex," IEEE Transactions on Image Processing, 6(3), Mar., 1997. https://doi.org/10.1109/83.557356

[9] R. Chouhan, C. Pradeep Kumar, R. Kumar, and R. K. Jha, "Contrast Enhancement of Dark Images using Stochastic Resonance in Wavelet Domain," International Journal of Machine Learning and Computing, 2(5), 2012. http://doi.org/10.7763/IJMLC.2012.V2.220.

[10] X. Fu, Y. Liao, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A Probabilistic Method for Image Enhancement With Simultaneous Illumination and Reflectance Estimation," in IEEE Transactions on Image Processing, 24(12), pp. 4965-4977, Dec., 2015. https://doi.org/10.1109/TIP.2015.2474701

[11] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A New Low-Light Image Enhancement Algorithm Using Camera Response Model," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, pp. 3015-3022, 2017. http://doi.org 10.1109/ICCVW.2017.356

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, 86(11), pp. 2278-2324, Nov. 1998. http://doi.org/10.1109/5.726791

[13] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), pp. 640-651, 1 April 2017. http:// 10.1109/TPAMI.2016.2572683

[14] H. Noh, S. Hong and, B. Han, "Learning Deconvolution Network for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 1520-1528, 2015. http://doi.org/10.1109/ICCV.2015.178

[15] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the Exclusively Dark dataset," Computer Vision and Image Understanding, 178, pp 30-42, 2019. https://doi.org/10.1016/j.cviu.2018.10.010.

[16] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and John Paisley, "A fusion-based enhancing method for weakly illuminated images," Signal Processing, 129, 2016. https://doi.org/10.1016/j.sigpro.2016.05.031.

[17] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Enhancing by saliency-guided decolorization," CVPR 2011, Colorado Springs, CO, USA, pp. 257-264, 2011. http://doi.org/10.1109/CVPR.2011.5995414

[18] C. Liu and T. Liu, "A sparse linear model for saliency-guided decolorization," 2013 IEEE International Conference on Image Processing, Melbourne, VIC, pp. 1105-1109, 2013. http:// doi.org/10.1109/ICIP.2013.6738228

[19] K. He, J. Sun, and X. Tang, "Guided Image Filtering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(6), pp. 1397-1409, June 2013. http://doi.org/10.1109/TPAMI.2012.213

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 165-170 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Photodecoloration of Methyl Orange Solution Assisted by ZrS_3 Powders

Sofya Artemkina*[1,2], Anastassiia Poltarak[1,2], Pavel Poltarak[1,2], Igor Asanov[1], Vladimir Fedorov[1,2]

[1]*Nikolaev Institute of Inorganic Chemistry SB RAS, Novosibirsk 630090, Russian Federation*

[2]*Novosibirsk State University, Department of Natural Sciences, Novosibirsk 630090, Russian Federation*

ARTICLE INFO

Article history:
Received: 20 November, 2018
Accepted: 01 February, 2019
Online : 07 February, 2019

Keywords:
Methyl orange
Decoloration
Zirconium trisulfide
Synthesis
Morphology
Catalysis

ABSTRACT

Zirconium trisulfide ZrS_3 is a representative of transition metal polysulfides containing sulfur as S^{-1} in polysulfide, usually disulfide S_2^{2-} groups. Semiconductive zirconium trisulfide which absorbs visible light near UV edge was considered as a possible photocatalyst. We experimentally studied photodecoloration of methyl orange in presence of ZrS_3. It was shown for the first time that crystalline ZrS_3 strongly deepens photodegradation reaction, and in one case the methylene orange conversion reached almost 100%. The rates of degradation curves were associated with the ZrS_3 samples morphology; the best result revealed for microribbons ZrS_3 synthesized at 650°C.

1. Introduction

Transition metal sulfides are well-known class of inorganic compounds. The very popular representatives CdS, MoS_2 and some others are considered as materials for luminescence [1], photocatalysis [2], piezoelectric materials [3], anodic components for lithium batteries [4], etc. [5].

The distinctive feature of transition metal polysulfides (TMP) is a polysulfide group which is usually a disulfide group $(S_2)^{2-}$ coordinated to metal atoms. For instance, crystalline VS_4 [6], NbS_3 [7], TiS_3 [8], as well as amorphous MoS_3 [9], MoS_4 and WS_5 [10] include disulfide $(S_2)^{2-}$ groups along with sulfide anions S^{2-} coordinated to metal atoms. Like layered transition metal disulfides, TMP also possess low-dimensional structures (layered – crystalline trisulfides, chained – crystalline VS_4, amorphous CrS_3, MoS_3, etc. – presumably chained).

Zirconium trisulfide ZrS_3 is a typical representative of layered transition metal trisulfides. The crystal structure of ZrS_3 (attributed to $ZrSe_3$ structural type [8]) (Fig. 1) comprise double layers whose surfaces are produced by disulfide groups $(S_2)^{2-}$. Particularly, a basic structural fragment prism $\{ZrS_{6/2}\}$, is constructed from two disulfide $(S–S)^{2-}$ groups as well as two sulfide groups S^{2-}, the Zr atom is situated near the center of the prism (Fig. 1). Such prisms $\{ZrS_{6/2}\}$ with metal atoms situated close to the prism centers are connected to each other via common triangle bases to form infinite columns which are oriented along *b* axis. Additionally, these columns are connected to the neighboring ones forming layers two prisms thick. The infinite layers are bind to the neighboring *via* van der Waals S…S contacts. Hence the layered zirconium trisulfide may be described by ionic model $Zr^{4+}(S_2)^{2-}S^{2-}$.

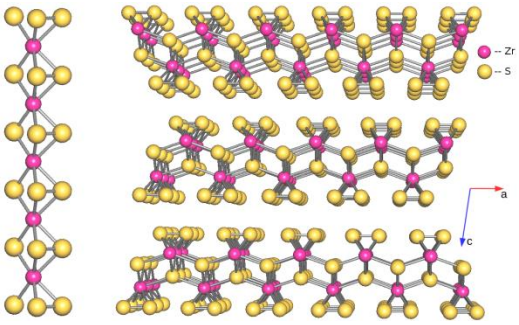


Figure 1: Structure of the wedge-shaped columns (left) within one polymeric layer and arrangement of the layers (right) in the crystal structure of triclinic ZrS_3

Presence of the disulfide group introduces new chemical properties which could be useful in known physico-chemical processes required in modern industry. Sulfur in the disulfide group is charged as –1, so it can be oxidized or reduced [11] in 1-electron ox-red processes, so these systems can be considered as

*Artemkina Sofya, NIIC SB RAS, 3, Akad. Lavrentiev Ave., 630090 Novosibirsk, Russia, +7 383 330 92 53, artem@niic.nsc.ru

electron reservoirs [12]. This property can meet an application in high capacity alkaline batteries [4], [13]. In the ionic model approximation, trisulfides MS_3 (M = Ti, Zr, Hf) are d^0-complexes, a diamagnetic semiconductors. The semiconductor band gaps of the trisulfides MS_3 (M = Ti, Zr, Hf) were optically defined (1.8, 2.0, and 2.2 eV, respectively), and were recently examined as cathodes in hydrogen evaluation reaction [14]. In addition, ZrS_3 nanomaterial was used as efficient anode for catalytic oxygen evaluation reaction in both alkaline and neutral aqueous solutions [15].

Photocatalytic degradation of organic molecules is an actual theme now, and it becomes more popular once renewable energy source is used (solar irradiation); the nature of substrates is dictated by polluting agents in natural waters and soils. Semiconductor photocatalysis is an important option for solar energy conversion and pollutant degradation according to the idea of green chemistry. For purposes of degradation of the pollutants, a known set of catalysts is already designed [16]: double layered hydroxides [17], titanium dioxide [18], composites based on TiO_2 [19], other oxides (ZnO [20], WO_3 [21]), composites with oxides [22-24, 25, 26], as well as composites with MoS_2 [27-29].

Many transition metal di- and polysulfides are dark or metal black in color. A pair of "colored exceptions" can be distinguished here, ZrS_3 and HfS_3, which are orange and dark red, respectively. According to study [30], these compounds absorb visible light in near UV region, so we decided to examine ZrS_3 compound as photocatalyst in a model reaction of methyl orange degradation. Shifting from transition metal disulfides (MoS_2, WS_2) to transition metal trisulfides as photocatalyst is something new in the present work.

2. Experimental

2.1. Materials

Zirconium powder 99% and crystalline sulfur 99.9% (Acros), were used for synthesis of ZrS_3 samples; TiO_2 Degussa P25 (Degussa AG, Germany), methyl orange (Reakhim, Russia) and distilled water were used for carrying out photodecolorization experiments.

2.2. Syntheses of zirconium trisulfide at different temperatures

ZrS_3 was obtained by heating a stoichiometric mixture of Zr and S. Mixtures of powder zirconium (4.868 g 0.05336 mol) and crystalline sulfur (5.132 g 0.1601 mol) were evacuated in three quartz ampules (about 20 ml volume) and sealed. The syntheses were carried out at temperatures 350, 500, and 650°C for 150 hours (up to these temperatures, the ampoules were heated for 15 h). The ampules were cooled down with the furnace. The products were thin powders different shades of red-orange. The products were heated in dynamic vacuum at 200°C during 1 h in order to remove elementary sulfur after syntheses.

2.3. Methods and apparatus

X-ray powder diffraction patterns for solids including exfoliated samples were collected with a Philips PW 1830/1710 automated diffractometer (Cu K_α radiation, graphite monochromator, silicon plate as an external standard). Raman spectra were recorded with a LabRam HR Revolution (Horiba Scientific) instrument at wavelength 488nm. UV spectra were recorded with an Agilent Cary 60 Spectrophotometer in the range of 200 – 800 nm. Eppendorf Centrifuge 5430 equipped with container for 15 and 50 ml tubes used for centrifugation of the reactive mixtures. Scanning electron microscopy (SEM) images were collected with Hitachi S3400N instrument.

X-ray photoelectron spectroscopy (XPS) measurements were carried out on a Specs Phoibos-150 spectrometer with an Al Kα monochromatic excitation. The pass energy of an electron analyzer was set at 20 eV. For the compensation of a charging effect from non-conductive samples, a low energy electron beam was applied. Binding energies were measured from the C1s level (285.0 eV) of surface hydrocarbon contaminations. Relative atomic concentrations are calculated from the measured areas of spectra taking into account photoionization cross-sections, inelastic mean free paths and the transmission function of the spectrometer.

The Brunauer–Emmett–Teller (BET) surface areas were determined by N_2 adsorption 5-point measurements in pressure range p/p_0 = 0.05–0.25 using SORBTOMETER-M surface area analyzer. The ZrS_3 powder samples were preliminary heated at 200°C during 1 hour in nitrogen flow.

2.4. Photodecoloration experiment

The photocatalytic activity of ZrS_3 powders was evaluated by standard decoloration reaction of dye methyl orange (MO). The reactor for the decoloration experiments consisted of a cylindrical beaker on a magnetic stirrer, with a common LH9-U/BLB/G23 black light lamp (Camelion). Experimental conditions for methyl orange (MO) decoloration on powders were: aqueous solution of MO with initial concentration 0.022 mM, amount of ZrS_3 powder 0.175 g/L (14 mg per each experiment), and UV irradiation time was kept constant for 2-3 hours. The aqueous suspensions were stirred throughout the experiment. In the experiment, to 80 ml of aqueous solution of MO (C = 0.022 mM) 14 mg of ZrS_3 powder was added, and mixed during 3 hours excluding irradiation of any light (for dark sorption). At the stage of dark sorption several points were documented in the following way: 1 ml of suspension was collected from the solution with Pasteur pipette, centrifuged with acceleration 3000 *g* for 10 min for removal of ZrS_3 solid, the transparent solution decanted, and the UV-Vis spectrum in region 200-800 nm written for the solution. After 3 hours of dark sorption a lamp was turned on above the tested mixture and every 15 min points were examined in the same way. Three series of experiments were carried out, each series three times: using powder samples of ZrS_3 synthesized at 350 (ZrS_3-350), 500 (ZrS_3-500), and 650°C (ZrS_3-650). Graphics of MO absorbance *vs* time of irradiation were built for comparison.

2.5. Photodecoloration experiment with Degussa P25

The experiment with Degussa P25® was carried out in similar way, using the same experimental setup. The portion of powder Degussa P25 was 14 mg per every degradation experiment.

After the experiments the photocatalysts were washed with 20 ml of water, centrifuged and decanted, and then washed with 20 ml of ethanol. After these operations the solids were dried in

air. They were examined by XRD and Raman spectroscopy, as well as XPS for ZrS_3-650.

3. Results and discussion

3.1. Initial powder ZrS_3 samples

Zirconium trisulfide samples were synthesized from the elements in evacuated quartz ampules at different temperatures according to reaction which is rather common for synthesis of transition metal trichalcogenides [31, 32]:

$$Zr + 3S \rightarrow ZrS_3 \quad (T = 350, 500, 650°C).$$

The yields were quantitative or near-quantitative. Preliminary EDS tests revealed a quantity of elementary sulfur in the products, so we heated the ZrS_3 samples in dynamic vacuum in order to remove elementary sulfur. The products were thin powders, red-orange in color (red for ZrS_3-350, yellow-orange for ZrS_3-500 and ZrS_3-650). Powder diffraction patterns (XRD) revealed only monoclinic ZrS_3 phase (ICSD 42-073) as crystalline product of reaction in the cases of temperatures 500 and 650°C (Fig. 2). Positions of some observable reflections were as follows: *2θ* 9.97° (*001*), 20.01° (*002*), 35.45° (*200*), 40.55° (*004*), 43.59° (*211*), 50.42° (*020*), 62.99° (*006*). Small quantity of unreacted metal zirconium was found in ZrS_3-350 (2θ 36.6°). Coherent scattering domain size (CSDS) calculated by Scherrer formula [33] (applied to *001* reflection) from the powder patterns gave 14, 29, and 77 Å for ZrS_3-350, ZrS_3-500, and ZrS_3-650, respectively. So, we noticed here that rise in temperature of synthesis enhances the crystalline ordering of the samples ZrS_3. The Raman spectra of the ZrS_3 samples were very close to each other, and the vibration bands A_g 148 cm^{-1} (150 cm^{-1} from [34]), 276 (280), 317 (322), 525 (530) cm^{-1} were agree with the literature data, the 525 (530) cm^{-1} band testifies presence of $(S–S)^{2–}$ groups in the samples.

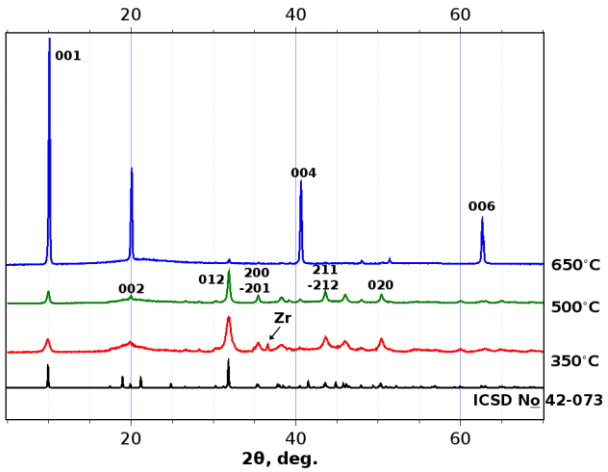


Figure 2: X-ray powder diffraction patterns for crystalline samples ZrS_3-350 (red line), ZrS_3-500 (green line), and ZrS_3-650°C (blue line), and theory calculated XRD pattern for ZrS_3 (black line)

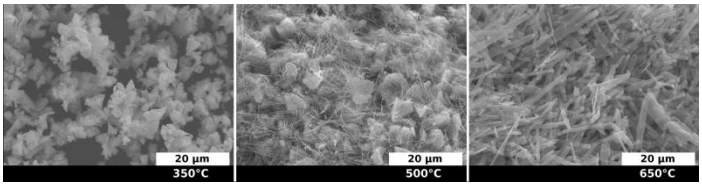
EDS data for the prepared ZrS_3 samples showed rather low content of sulfur (Zr : S ratios were from 2.2 to 2.4), so we considered that the materials demonstrated mainly surface stoichiometry which could include some deal of oxides also. These data contradicted with the XRD data where clear crystalline monoclinic ZrS_3 phase occurred. In order to define the exact Zr : S ratios in these samples, we carried out gravimetric experiments: carefully weighed ZrS_3 samples were put into pre-weighed crucibles and heated in a muffle furnace at 600°C during 20 hours. The products of calcination in the crucibles were weighed, and the quantity of sulfur was calculated using mass difference according to the equation:

$$ZrS_x + (1+x)O_2 = ZrO_2 + xSO_2.$$

After weighing the calcination products, they were examined by X-ray powder diffraction in order to make sure that ZrO_2 (gravimetric form) were the only calcination product. It was found that the bulk material ZrS_3-350 was $ZrS_{2.56}$, ZrS_3-500 was $ZrS_{2.76}$, and ZrS_3-650 was $ZrS_{2.76}$. So we concluded from this data that the ZrS_3 samples synthesized from stoichiometric mixtures of Zr and S are monoclinic ZrS_3 albeit possess remarkable deficiency in sulfur.

Probably our ZrS_3 samples have such a sulfur deficiency due to synthetic procedure: only 3 equivalents of S were taken for reaction per one equivalent of Zr. Crystals of zirconium trisulfide with formulae very close to ZrS_3 are synthesized in gas-transport reactions with some excess of sulfur [35], longer heating, or the stoichiometric mixture which was heated at higher temperature (1000°C) [35], or applied transport species (for example, bromine) [36].

SEM images of the powder ZrS_3 samples revealed different morphologies (Fig. 3). ZrS_3-350 sample is aggregates of lamellar slabs about 5μm size; ZrS_3-500 sample is a mixture of particles of two types: aggregates of lamellar slabs ca. 5μm size and fibers up to 20 μm long; ZrS_3-650 sample is ribbons 0.5-1 μm width and ca. 20μm long. Visible texturing is observed for ZrS_3-650 sample (Fig. 2): apparently, the preferred orientation of the ribbons is layering in parallel to their wide side.



a) b) c)

Figure 3: SEM images of powder samples synthesized at ZrS_3-350 (a), ZrS_3-500 (b), and ZrS_3-650 (c)

Strong dependence of the BET surface areas on the ZrS_3 morphology (min BET 9 m^2/g for ZrS3-650, Table 1) indicates that the lamellar slab aggregates make the main contribution to the surface area: BET of the samples ZrS_3-350 and ZrS_3-500 having components of such morphology, are noticeably higher (49 and 24 m^2/g, respectively). Degree of dark sorption correlated with the surface area values: the percentage of absorbed MO on ZrS_3 powders increased while increasing their BET in series ZrS_3-650 < ZrS_3-500 < ZrS_3-350. The minimal dark adsorption of ZrS_3-650 is in agreement with the maximal CSDS value, 77 Å, and the minimal BET surface.

3.2. Photodecoloration of MO in presence of ZrS_3 powders

ZrS_3 powders were used as photocatalysts for decomposing of dye molecule MO for the first time (Fig. 4). ZrS_3 powders were used as synthesized. Photocatalytic activity of ZrS_3 was evaluated

in terms of the degradation of MO aqueous solution under UV irradiation. The efficiency of the dye degradation was calculated based on the changes in absorption of MO in UV-Vis spectra. At the lamp irradiation, MO didn't decolorate in absence of ZrS_3.

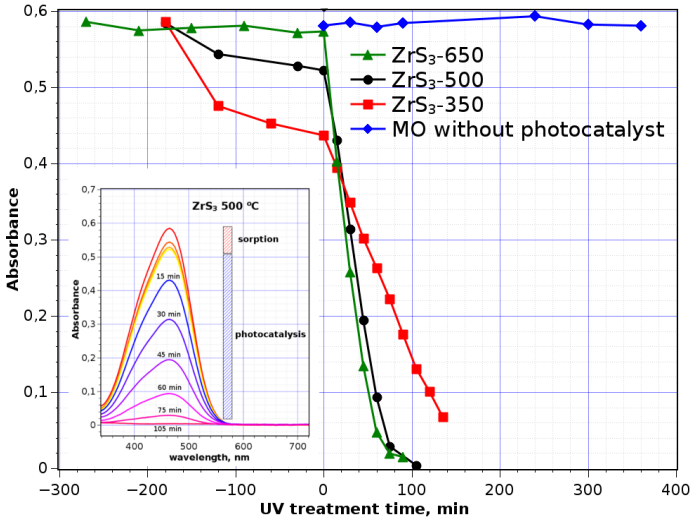


Figure 4: Dependence of MO absorbance in the presence of ZrS_3 *vs* time of UV irradiation exposure. Systematic errors of measurement are from minimum three experiments for each ZrS_3 sample

From the UV-Vis spectroscopy data (Fig. 4, inset) one can see that MO absorbs on ZrS_3 particles in the cases ZrS_3-350 and ZrS_3-500 in the dark phase whereas presence of ZrS_3-650 decreases MO concentration to a lesser degree. After start of the irradiation, the concentration of MO in the reaction mixture was decreased noticeably, that indicated start of MO photodegradation.

The decoloration dependencies vs time of irradiation are averaged from at least three experiments, and the standard deviations of the data points not exceed 5% from absorbance values at time starting UV treatment (t = 0 min).

Table 1: Some parameters of the decoloration process in presence of different ZrS_3 crystalline samples

Crystalline ZrS_3 sample	ZrS_3-350°C	ZrS_3-500°C	ZrS_3-650°C
Coherent scattering domain size (CSDS), Å	14	29	77
BET surface area, m^2/g	49	24	9
Adsorption of MO in "dark" time, %	26	11	2.2
Time of irradiation at 50% degradation of MO (0.022mM), min	48	36	27
Time of irradiation at 90% degradation of MO (0.022mM), min	144	71	59
Time of irradiation at 50% degradation of MO (0.022mM) in presence of TiO_2 Degussa P25, min		138	

At the initial concentration of MO in aqueous solution 0.022 M, photodegradation time was confined by two hours. The longest MO degradation process concerns to ZrS_3-350 sample, 140 min with 90% of the initial MO amount. The photodegradation dependences have linear character at least for the first 60 min, for ZrS_3-500 and ZrS_3-650 they are very close while the line for ZrS_3-350 has gentler slope, and doesn't demonstrate complete degradation of MO in 140 min of the experiment.

One can see ZrS_3-650 is the most efficient sample from the three. The reason of this may be the minimal dark sorption, so the sample surface is better accessible for dissolved oxygen; ZrS_3-650 also possesses larger size of coherent scattering domain (77Å) that allows hoping to have more regular surface of the particles. The MO degradation results obtained are comparable with MO degradation results in the presence of MoS_2 of two morphologies: nanorods and nanoflowers [37].

3.3. Characterization of ZrS_3 powders after the photodecoloration experiments

After the photodegradation experiments, ZrS_3 powder samples retained their structural identity (XRD for ZrS_3-650 is presented in Fig. 5). None of crystalline ZrO_2 phases were detected in the XRD patterns. Raman spectra of ZrS_3 after the decoloration experiments revealed ZrS_3 spectra which were very similar to each other: 105, 119, 146, 239, 275, 316, 357, 523 cm^{-1}. These bands are in good accordance with the literature data: 107, 120, 148, 241, 277, 317, 355, 525 cm^{-1} [38]. We didn't find in the Raman spectra band respected to amorphous ZrO_2 (148, 263, 476 cm^{-1}) as in [39].

The analysis of XPS data before and after the photocatalytic experiments for ZrS_3-650 as critical sample, revealed certain differences. In the initial powder sample ZrS_3-650 (Fig. 6), the atomic concentration ratio S/Zr was 2.78. S2p spectrum contained two doublets from S_2^{2-} and S^{2-} with positions of $S2p_{3/2}$ components at 162.6 and 161.6 eV, respectively; the ratio of their intensities was 1.78. The values of bond energies were well agreed with the literature data [40]. Spectrum $Zr3d_{5/2, 3/2}$ also contains two doublets with spin-orbital splitting 2.4 eV. The $Zr3d_{5/2}$ components were situated at 181.2 and 182.9 eV with intensity ratio 9:1. The first component belonged to zirconium in ZrS_3 while the second might be associated with an oxide phase on surface of the ZrS_3 particles [41]. The composition of the surface sulfide phase may be thereby formulated as $Zr_{0.9}(S_2)_{0.89}S$.

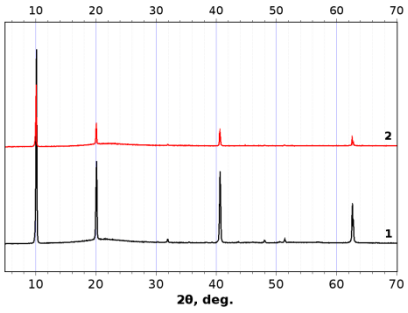


Figure 5: XRD patterns of ZrS_3-650 before (line 1) and after (line 2) photodegradation MO experiment

In the sample ZrS_3-650 after the photodegradation experiment, weak peaks in the region 168.0 eV appeared in $S2p_{3/2}$ spectrum that corresponds to sulfate groups on the surface. Component

corresponding to an oxide phase, strongly arose in the $Zr3d_{5/2, 3/2}$ spectrum. These data indicate the oxidation of the ZrS_3 particles surface after photocatalysis experiment.

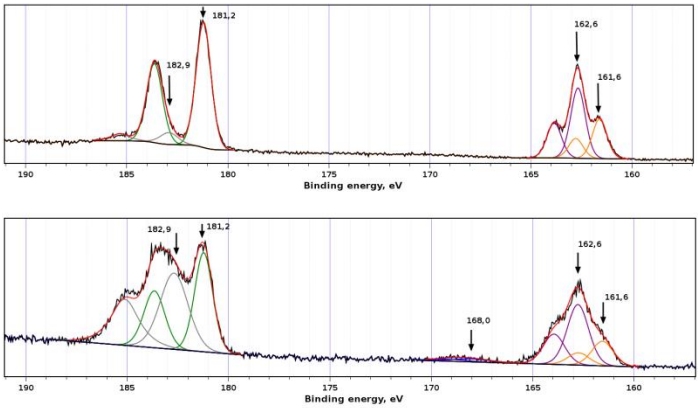


Figure 6: XPS data for ZrS_3-650 powder sample, as-synthesized (upper graphic) and after photocatalysis experiment (lower graphic)

3.4. Comparison with Degussa P25

We carried out photocatalytic experiment with Degussa P25 using the same experimental setup in order to compare photocatalytic ability of ZrS_3 with the international reference TiO_2 (TiO_2, 80% anatase and 20% rutile). 14 mg of Degussa P25 in 80 ml of 0.022 mM MO solution was used for the experiment. As a result of experiment in our setup, MO decolorated in presence of Degussa P25 at 50% in 138 min. Although kinetics of MO decoloration was faster for ZrS_3-650 than for Degussa P25 here, we discussed here only the first cycle of decoloration. Additional studies are required to test ZrS_3 powders in series of decoloration cycles, and examination of the catalysts after the experiments.

The preliminary experiment on photodegradation of MO in aqueous solution in presence of powder ZrS_3 opens way to design of new light-sensitive materials for photocatalysis and other applications. Some additional studies are to carry out in order to realize the mechanism and other characteristics of the photodegradation observed. Authors believe that photosensitive properties of ZrS_3 (and probably some another transition metal trisulfides) may be enhanced by different ways (crystal or surface modification, assembly in composites) to obtain new materials for photocatalysis.

4. Conclusion

Zirconium trisulfide revealed appreciable photocatalytic ability in dye molecules decoloration, on the example of methyl orange decoloration for the first time. The surface of the photocatalyst increases the content of oxide phase after the degradation reaction. It was ascertained that the photocatalytic behavior of ZrS_3 is dependent on its morphology, which in turn is dependent on temperature of synthesis of ZrS_3. Based on qualitative level results, ZrS_3-containing species are considered as perspective materials for photocatalysis.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was financially supported by Russian Science Foundation (grant RSCF 13-14-00674). Authors are grateful to Dr. Eugeny Maksimovsky for collecting of SEM images and Dr. Konstantin Kovalenko for BET analysis.

References

[1] P.F. Smet, I. Moreels, Z. Hens, and D. Poelman, "Luminescence in Sulfides: A Rich History and a Bright Future" Materials, 3 (4), 2834, 2010. https://doi.org/10.3390/ma3042834

[2] J. Mao, Y. Wang, Z. Zheng, and D. Deng, "The rise of two-dimensional MoS2 for catalysis" Frontiers of Physics, 13 (4), 138118, 2018. https://doi.org/10.1007/s11467-018-0812-0

[3] K. Jinzong, L. Yudong, Z. Yaxing, and Z. Junyi, "Progress in piezotronics of transition-metal dichalcogenides" Journal of Physics D: Applied Physics, 51 (49), 493002, 2018. https://doi.org/ 10.1088/1361-6463/aadb15

[4] E.D. Grayfer, E.M. Pazhetnov, M.N. Kozlova, S.B. Artemkina, and V.E. Fedorov, "Anionic redox chemistry in polysulfide electrode materials for rechargeable batteries" ChemSusChem, 10 (24), 4805–4811, 2017. https://doi.org/10.1002/cssc.201701709

[5] C. Yan, C. Gong, P. Wangyang, J. Chu, K. Hu, C. Li, X. Wang, X. Du, T. Zhai, Y. Li, and J. Xiong, "2D Group IVB Transition Metal Dichalcogenides" Advanced Functional Materials, 28 (39), 1803305, 2018. https://doi.org/10.1002/adfm.201803305

[6] M.N. Kozlova, Y.V. Mironov, E.D. Grayfer, A.I. Smolentsev, V.I. Zaikovskii, N.A. Nebogatikova, T.Y. Podlipskaya, and V.E. Fedorov, "Synthesis, Crystal Structure, and Colloidal Dispersions of Vanadium Tetrasulfide (VS_4)" Chemistry – A European Journal, 21 (12), 4639-4645, 2015. https://doi.org/10.1002/chem.201406428

[7] J. Rijnsdorp, and F. Jellinek, "The crystal structure of niobium trisulfide, NbS_3" Journal of Solid State Chemistry, 25 325-328, 1978.

[8] S. Furuseth, L. Brattas, and A. Kjekshus, "On the crystal structures of TiS_3, ZrS_3, $ZrSe_3$, $ZrTe_3$, HfS_3, and $HfSe_3$" Acta Chemica Scandinavica A, 29 623-631, 1975.

[9] S.J. Hibble, and G.B. Wood, "Modeling the Structure of Amorphous MoS3: A Neutron Diffraction and Reverse Monte Carlo Study" Journal of the American Chemical Society, 126 (3), 959-965, 2004. https://doi.org/10.1021/ja037666o

[10] D.A. Rice, S.J. Hibble, M.J. Almond, K.A.H. Mohammad, and S.P. Pearse, "Novel low-temperature route to known (MnS and FeS_2) and new (CrS_3, MoS_4 and WS_5) transition-metal sulfides" Journal of Materials Chemistry, 2 (8), 895-896, 1992. https://doi.org/10.1039/JM9920200895

[11] M.N. Kozlova, E.D. Grayfer, P.A. Poltarak, S.B. Artemkina, A.G. Cherkov, L.S. Kibis, A.I. Boronin, and V.E. Fedorov, "Oxidizing Properties of the Polysulfide Surfaces of Patronite VS_4 and NbS_3 Induced by (S2)2- Groups: Unusual Formation of Ag_2S Nanoparticles" Advanced Materials Interfaces, 1700999, 2017. DOI: 10.1002/admi.201700999

[12] J. Rouxel, "Anion–Cation Redox Competition and the Formation of New Compounds in Highly Covalent Systems" Chemistry – A European Journal, 2 (9), 1053-1059, 1996. https://doi.org/10.1002/chem.19960020904

[13] V.V.T. Doan-Nguyen, K.S. Subrahmanyam, M.M. Butala, J.A. Gerbec, S.M. Islam, K.N. Kanipe, C.E. Wilson, M. Balasubramanian, K.M. Wiaderek, O.J. Borkiewicz, K.W. Chapman, P.J. Chupas, M. Moskovits, B.S. Dunn, M.G. Kanatzidis, and R. Seshadri, "Molybdenum Polysulfide Chalcogels as High-Capacity, Anion-Redox-Driven Electrode Materials for Li-Ion Batteries" Chemistry of Materials, 28 (22), 8357-8365, 2016. https://doi.org/10.1021/acs.chemmater.6b03656

[14] E. Flores, J.R. Ares, I.J. Ferrer, and C. Sanchez, "Synthesis and characterization of a family of layered trichalcogenides for assisted hydrogen photogeneration" Physica Status Solidi, 10 (11), 802-806, 2016. https://doi.org/10.1002/pssr.20160016

[15] J. Xie, R. Wang, J. Bao, X. Zhang, H. Zhang, S. Li, and Y. Xie, "Zirconium Trisulfide Ultrathin Nanosheets as Efficient Catalyst for Water Oxidation in Both Alkaline and Neutral Solutions" Inorganic Chemistry Frontiers, 1 751-756, 2014. https://doi.org/10.1039/C4QI00127C

[16] C. Byrne, G. Subramanian, and S.C. Pillai, "Recent advances in photocatalysis for environmental applications" Journal of Environmental Chemical Engineering, 6 (3), 3531-3555, 2018. https://doi.org/10.1016/j.jece.2017.07.080

[17] G. Starukh, "Photocatalytically enhanced cationic dye removal with Zn-Al layered double hydroxides" Nanoscale Research Letters, 12 (1), 391, 2017.
[18] C. Chen, W. Ma, and J. Zhao, "Semiconductor-mediated photodegradation of pollutants under visible-light irradiation" Chemical Society Reviews, 39 (11), 4206-4219, 2010. https://doi.org/***
[19] L. Zhang, Y. He, Y. Wu, and T. Wu, "Photocatalytic degradation of RhB over $MgFe_2O_4/TiO_2$ composite materials" Materials Science and Engineering: B, 176 (18), 1497-1504, 2011. https://doi.org/10.1016/j.mseb.2011.09.022
[20] R. Kumar, G. Kumar, and A. Umar, "Zinc oxide nanomaterials for photocatalytic degradation of methyl orange: a review" Nanoscience and Nanotechnology Letters, 6 (8), 631-650, 2014. https://doi.org/10.1166/nnl.2014.1879
[21] P. Dong, G. Hou, X. Xi, R. Shao, and F. Dong, "WO_3-based photocatalysts: morphology control, activity enhancement and multifunctional applications" Environmental Science: Nano, 4 (3), 539-557, 2017. https://doi.org/10.1039/C6EN00478D
[22] X. Yang, H. Fu, K. Wong, X. Jiang, and A. Yu, "Hybrid Ag@TiO_2 core–shell nanostructures with highly enhanced photocatalytic performance" Nanotechnology, 24 (41), 415601, 2013. https://doi.org/10.1088/0957-4484/24/41/415601
[23] J. Luo, X. Zhou, L. Ma, and X. Xu, "Enhanced visible-light-driven photocatalytic activity of WO_3/BiOI heterojunction photocatalysts" Journal of Molecular Catalysis A: Chemical, 410 168-176, 2015. https://doi.org/10.1016/j.molcata.2015.09.019
[24] S. Chen, Y. Hu, X. Jiang, S. Meng, and X. Fu, "Fabrication and characterization of novel Z-scheme photocatalyst WO_3/g-C_3N_4 with high efficient visible light photocatalytic activity" Materials Chemistry and Physics, 149-150 512-521, 2015. https://doi.org/10.1016/j.matchemphys.2014.11.001
[25] S. Bai, K. Zhang, J. Sun, R. Luo, D. Li, and A. Chen, "Surface decoration of WO_3 architectures with Fe_2O_3 nanoparticles for visible-light-driven photocatalysis" CrystEngComm, 16 3289-3295, 2014. https://doi.org/10.1039/C3CE42410C
[26] S.-M. Lam, J.-C. Sin, A.Z. Abdullah, and A.R. Mohamed, "Sunlight responsive WO_3/ZnO nanorods for photocatalytic degradation and mineralization of chlorinated phenoxyacetic acid herbicides in water" Journal of Colloid and Interface Science, 450 34-44, 2015. https://doi.org/10.1016/j.jcis.2015.02.075
[27] L. Shi, Z. He, and S. Liu, "MoS_2 quantum dots embedded in g-C_3N_4 frameworks: A hybrid 0D-2D heterojunction as an efficient visible-light driven photocatalyst" Applied Surface Science, 457 30-40, 2018. https://doi.org/10.1016/j.apsusc.2018.06.132
[28] Q. Qin, Q. Shi, W. Ding, J. Wan, and Z. Hu, "Efficient hydrogen evolution and rapid degradation of organic pollutants by robust catalysts of MoS_2/TNT@CNTs" International Journal of Hydrogen Energy, 43 (33), 16024-16037, 2018. https://doi.org/10.1016/j.ijhydene.2018.07.051
[29] H. Li, H. Shen, L. Duan, R. Liu, Q. Li, Q. Zhang, and X. Zhao, "Enhanced photocatalytic activity and synthesis of ZnO nanorods/MoS2 composites" Superlattices and Microstructures, 117 336-341, 2018. https://doi.org/10.1016/j.spmi.2018.03.028
[30] S. Kurita, M. Tanaka, and F. Levy, "Optical spectra near band edge of ZrS_3 and $ZrSe_3$" Phsical Review B, 48 (3), 1356-1360, 1993. https://doi.org/10.1103/physrevb.48.1356
[31] A.A. Stabile, L. Whittaker, T.L. Wu, P.M. Marley, S. Banerjee, and G. Sambandamurthy, "Synthesis, characterization, and finite size effects on electrical transport of nanoribbons of the charge density wave conductor $NbSe_3$" Nanotechnology, 22 485201, 2011. https://doi.org/10.1088/0957-4484/22/48/485201
[32] K. Endo, H. Ihara, K. Watanabe, and S.I. Gonda, "XPS study on valence band structures of transition-metal trisulfides, TiS_3, NbS_3, TaS_3" Journal of solid state chemistry, 39 215-218, 1981. https://doi.org/10.1016/0022-4596(81)90334-0
[33] A.L. Patterson, "The Scherrer formula for X-ray particle size determination" Physical Review, 56 978-982, 1939.
[34] H. Jin, D. Cheng, J. Li, X. Cao, B. Li, X. Wang, X. Liu, and X. Zhao, "Facile synthesis of zirconium trisulfide and hafnium trisulfide nanobelts: Growth mechanism and Raman spectroscopy" Solid State Sciences, 13 (5), 1166-1171, 2011. https://doi.org/10.1016/j.solidstatesciences.2010.12.017
[35] K. Giagloglou, J.L. Payne, C. Crouch, R.K. Gover, P.A. Connor, and J.T. Irvine, "Zirconium trisulfide as a promising cathode material for Li primary thermal batteries" Journal of the Electrochemical Society, 163 (14), A3126-A3130, 2016. https://doi.org/10.1149/2.1351614jes
[36] P. Gard, F. Cruege, C. Sourisseau, and O. Gorochov, "Single-crystal micro-Raman studies of ZrS_3, TiS_3, and several $Zr_{1-x}Ti_xS_3$ compounds ($0<x<=0.33$)" Journal of Raman Spectroscopy, 17 283-288, 1986. https://doi.org/10.1002/jrs.1250170310
[37] C. Wang, Y. Zhan, and Z. Wang, "TiO_2, MoS_2, and TiO_2/MoS_2 Heterostructures for Use in Organic Dyes Degradation" ChemistrySelect, 3 (6), 1713-1718, 2018. https://doi.org/10.1002/slct.201800054
[38] H. Jin, D. Cheng, J. Li, X. Cao, B. Li, X. Wang, X. Liu, and X. Zhao, "Facile synthesis of zirconium trisulfide and hafnium trisulfide nanobelts: Growth mechanism and Raman spectroscopy" Solid State Sciences, 13 1166-1171, 2011. https://doi.org/10.1016/j.solidstatesciences.2010.12.017
[39] V.G. Keramidas, and W.B. White, "Raman Scattering Study of the Crystallization and Phase Transformations of ZrO_2" J. Am. Ceram. Soc., 57 (1), 22-24, 1974. https://doi.org/10.1111/j.1151-2916.1974.tb11355.x
[40] F. Jellinek, R.A. Pollak, and M.W. Shafer, "X-ray photoelectron spectra and electronic structure of zirconium trisulfide and triselenide" Materials Research Bulletin, 9 (6), 845-856, 1974. http://dx.doi.org/10.1016/0025-5408(74)90121-4
[41] C. Morant, J.M. Sanz, L. Galán, L. Soriano, and F. Rueda, "An XPS study of the interaction of oxygen with zirconium" Surface Science, 218 (2), 331-345, 1989. https://doi.org/10.1016/0039-6028(89)90156-8

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 170-175 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Development & Implementation of Smart Vehicle Over Speeding Detector using IoT

Mohammad Ahmar Khan[1], Sarfraz Fayaz Khan [2], Taher Alawi Taher Barham BaOmar [1], Abdul Rahman Mohammed Aqeel Ba Omar[1]

[1]*Department of Management Information System, College of Commerce and Business Administration, Dhofar University, 211, Salalah, Sultanate of Oman*

[2]*Department of Information & Communication Technology, Algonquin College, K2G1V, Ottawa, Canada*

ARTICLE INFO

Article history:
Received: 06 January, 2019
Accepted: 28 February, 2019
Online: 25 March, 2019

Keywords:
Over speeding
Internet of Thing
Vehicle Over Speed Detection
Smart Vehicle

ABSTRACT

Nowadays over speeding is one of the most common traffic violations. Generally, over speeding is the result of restless and bad behavior of drivers. As the accident rates are increasing it is important to develop and implement a system which can automatically detect and report over speeding to the traffic control authorities as early as possible. Nearly all the roads are marked with speed limits depending upon the size of moving vehicles and heaviness of traffic, but some drivers habitually ignore this speed limit. The advancement in technology has replaced most of the manual or semi-automatic systems with an automated system. To add on to various systems in place, this research is making the use of Internet of Things to detect and report over speeding of the vehicle on which the device has been preinstalled. IoT is a technique to integrate various devices to exchange data among themselves. This research proposes the design, development and functioning of a smart device that helps in automatically detect and report to competitive authority, when so ever the subject vehicle exceeds the speed limit. The device has been developed based on the Global Positioning System (GPS) Technology using Raspberry Pi hardware and Android OS and has been practically tested on real time basis by installing it in a car.

1. Introduction

One of the main causes for accidents all over the world is rash driving. The traffic has increased in many folds but the system of monitoring the speed has been compromised. Rash driving causes severe risk not only to the driver & passenger but also to general public. Despite the fact that rash driving is a serious issue its present detection methods by patrol officers lacks efficiency [1].

Speed increase multiplies the hazard of accident and the risk of injury during the accidents [2]. At greater speed the vehicle needs a larger distance to stop. The need of hour is to develop a system that will detect and report the over speed of vehicles. The other factors like conditions of weather and driving at night also impact the manual system.

Daily death toll of India due to road accidents is greater than 4 times the yearly death toll from terrorism [3]. 1,39,671, people lost their lives on roads of India during 2014 that is 382 deaths every day. If 57,844 people lost their lives due to over speed nearly 6,969 Indians died because of speed driving post consumption of alcohol in 2014. The below figure shows the deaths caused due to accidents in roads

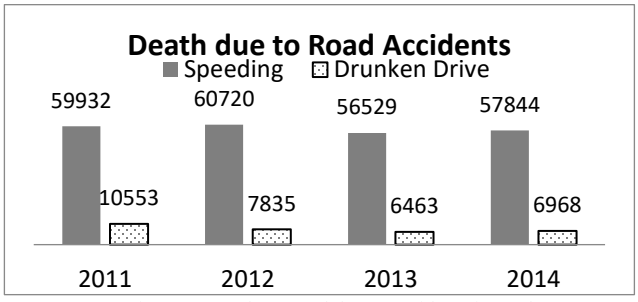


Figure 1: Deaths caused due to accidents in roads
Source: [3]

From the above figure it's evident that over speeding is the major cause of death on road accidents from 2011 to 2014. Over speed was responsible for 41 percent of deaths in accidents while dangerous or careless driving claimed 32 percent of deaths in accidents [4]. Mechanical defects in vehicles and bad weather conditions are other causes for deaths in accidents. Maharashtra and Tamil Nadu has reported several numbers of deaths due to over

*Dr. Mohammad Ahmar Khan, Dept. of MIS, CCBA, Dhofar University, PO Box 2509, PC 211, Sultanate of Oman +96898188243 mkhan@du.edu.om

speed of vehicles. The below figure shows the fatal road accidents by causes in 2015

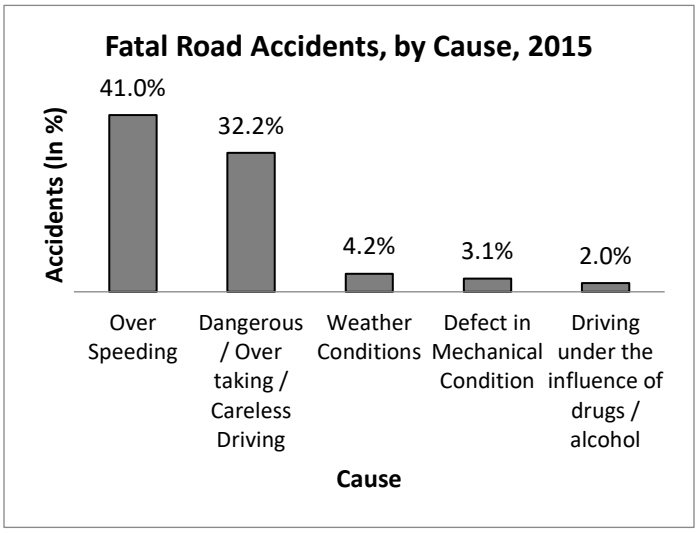


Figure 2: Fatal Road BY Cause in 2015
Source: National Crime Boards Bureau

In [5] the author has mentioned that a vehicle travelling on greater speed will have greater influence during the accident and may cause more number of injuries. Within the driver's fault category during 2016 accidents caused and killed individuals due to over speed accounted for greater number of 2,68,342 out of 4,03,599 accidents and 73,897 deaths out of 1,21,126 deaths respectively.

The author in [6] has stated that the over speed issue necessitates to develop a new way for managing the traffic system to avoid road accidents. This can be handled through real time traffic management using internet of things. Internet of things is a network which handles the physical devices through internet and this network offers rapid, exact and precise results. Internet of things system is designed to store entire database in PC and this storage is performed mainly through internet. Further this database is used accordingly as their applications and needs.

In [7] the author has mentioned that Internet of things system enhances the components to be accessed from far distanced which mainly reduces the work of humans or their interferences and this makes it much economical. Internet of things is linking various devices of application to each other through internet and this is possible because of sensors which support to transmit vast number of data location.

Internet of things connects the real world objects to virtual globe [8]. It comprises to a globe where living beings and physical objects as well as virtual environments and data, communicate with each other. Internet of Things based intelligent transportation systems are designed to assist the vision of Smart city which targets at using the powerful and advanced communications techniques for the administration of the citizens and the city. Internet of Things plays an essential part in the management of traffic enhancing the effectiveness of data transmission, developing the conditions of traffic, efficiency of management, reducing the costs of management and traffic safety.

Internet of Things is a global network linking the entire smart objects together [9]. The benefits of using internet of things in managing traffic are control of traffic, detection of human proximity, avoidance of theft, emergency response, avoidance of accident and autonomous vehicles. Similarly, the drawbacks of internet of things in managing traffic are network failure and security.

In [10] the authors had stated that enabling internet in every vehicle on the road can pave way for whole vehicle automation and traffic. The internet of things concept can be expanded to entire transport modes making an essential difference in the way that interaction exists between various modes of transport.

Thus it can be inferred that internet of things is an effective way for management of traffic and makes the road travel good for everybody in future.

2. Literature Review

The present modern human civilization era the traffic medium plays an essential part in the economic growth of the country [11]. Nowadays a control system is capable to manage such circumstances but not that much efficiently because it needs feedback from humans. A system is needed which must provide continuous update of satellite so that it can manage the traffic smoothly. This will be carried out using IoT using the transmission control protocol/internet protocol. A centralized server must be framed which many of the humans will view after entire analysis of traffic. Relying upon the traffic intensity the diversion path and waiting time will be determined.

In [12] the authors had presented a traffic and vehicle monitoring system based on Internet of Things. Traffic in urban regions and modern cities is making huge risk and is a main concern for the administration and public system. The circumstances namely traffic jams and accidents have become similar because of exponential development in vehicles on road. While human mistakes are one of the major reasons for these issues and lack of appropriate steps and adaptive control system of traffic is another reason. Security for vehicles is also essential. Even in this latent technical globe hackers are handling to break the aspects of security incorporated in modern vehicles. Several technologies namely Bluetooth, Radio Frequency Identification, GSM-GPS and Zigbee based systems were developed but they have restrictions in terms of usage and operation. IOT is a technology that links different objects and is developing at a rapid rate. This system is capable of resolving issues such as congestion in traffic, spotting of vehicle, early warnings considering traffic jams, emergency and VIP clearance of vehicles.

In [13], the researcher presented in their research that urban mobility is one of the unprecedented barriers to be handled in the management of a large city in the contemporary world. This research examines the every developing Urban population around the world and explains about the traffic systems in densely populated cities. Further an advanced system of traffic management is proposed and implemented using IoT (internet of things). The system is assisted by a circuit embedded in the vehicle which runs using radio frequency identification with clustered systems. The system functionalities involves effective light control of traffic, identification of parking space and security mechanism

of antitheft. The suggested architecture and process with big data analytics including Hadoop is presented and supervised learning methods are suggested that would support in deciding the roads standard, evaluating overall flow of traffic and examining the vehicles traffic path.

The authors in [14] proposed an intelligent traffic management based on internet of things. The congestion of traffic is a big problem in several big cities of India and the lights of traffic are use basically to manage the vehicle flow. Signal failure, huge number of vehicles, bad enforcement of law and poor management of traffic results in congestion of traffic. One of the main problems with the cities in India is that the existing infrastructure is not expanded much so there have only one choice available which manages traffic better. The traffic control system effectiveness relies on its capability to react on conditions of real time traffic. However conventional system of traffic control is not capable to perform this whatever the density of traffic low or high the signals are timed and performed according to that time. This result in increased congestion in traffic along the roads which outcome in pollution of air. This research proposed a smart traffic management system based on IoT to assist the traffic officers decision making.

In [15] the authors proposed a new method for managing the traffic system which utilizes an IoT (internet of things) concept. An intelligent traffic controller is framed with tools namely Pi-Camera, Raspberry Pi, IR sensors and Radio Frequency Identification. Raspberry Pi is the major tool which is employed to manage entire acts like a controller. The traffic density will be determined with the support of IR sensors and in order to provide zero traffic for emergency vehicles radio frequency identification technique is used. This study propose two modes of system control that is manual with introversion of human and automatic without any introversion of human.

In [16] the authors developed a smart traffic management system for control of traffic using automated electronic and mechanical devices. In the present smart city context particularly in the market and industrial zones the scenario of traffic is congested every time specifically at business hours peak time. Due to increasing development of vehicles and population in metropolitan and smart cities people faces huge issues at the main traffic points of business. It not only affects the delays in travelling but also contributes to pollution in environment as well as health risks due to pollution caused by fuels in vehicles. To resolve such problems several radiant urban communities are implementing smart control structures that run on traffic automation standards with hindrance of previously mentioned problems. In this research an enhanced traffic monitoring and control structure has been suggested that runs rapid transmission of information and their corresponding action. The simulation is undertaken using NS2 simulator which motivates the outputs in terms of good performance to manage the delay and hinder any accident due to profound congestion to a large extent.

In [17] the authors developed an internet of things based controller of traffic light in smart city. Nowadays the congestion of traffic has become a major issue in rapidly developing Indian cities which develops the air pollution. The vehicular density as well as consumption of fluid and this necessitates a new way for managing the traffic system. This has been handled through real time traffic density management using internet of things/ an intelligent traffic management system has been designed to manage the traffic system which involves components comprising of IR sensor, Raspberry Pi and display of LCD. Raspberry Pi is the major tool which is employed to manage the entire multitasking performance and IR sensor is employed to supervise the traffic density. The corresponding information is made feasible on site to show the status of traffic so that people will get early update and avoids traffic jams.

3. Design of the System

The study intends to propose, develop and install the Over Speeding Detector device for vehicle with the use of Internet of Things. The vehicles are fitted with over speeding detector device which has the capability for sharing information, recording and storing the data about the speed of vehicle. The following diagram shows the system architecture:

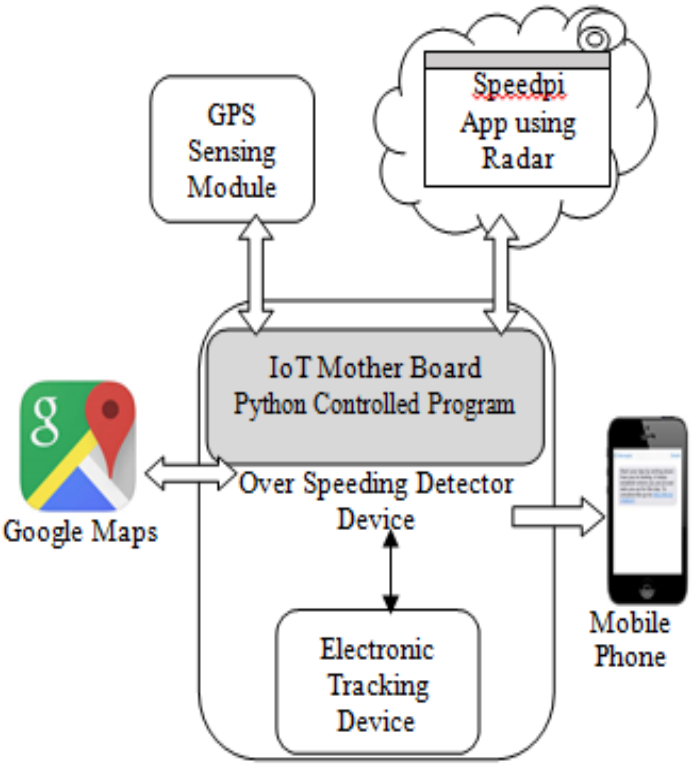


Figure 3: System Architecture
Source: Author

Google Maps:

The Google Maps Road Apps permits to plot GPS coordinates to geometry of the road and also to identify the vehicle's speed limits on the road segments. The following are services exposed for Google Maps Apps:

i. Snap to Roads: It returns the best-fit geometry of the road for a provided GPS coordinates set.
ii. Nearest Roads: It returns individual road divisions for a provided GPS coordinates set.
iii. Speed Limits: It returns the positioned speed limit for the road segment.

GPS Sensing Module:

GPS is a constellation of 24 or more satellites flying 20,350 km above the surface of the Earth. Each one circles the planet twice a day in one of six orbits to provide continuous, worldwide coverage.

GPS Working

i. GPS satellites broadcast radio signals providing their locations, status, and precise time (t1) from on-board atomic clocks.
ii. The GPS radio signals travel through space at the speed of light (c), more than 299,792 km/second.
iii. A GPS device receives the radio signals, noting their exact time of arrival (t2), and uses these to calculate its distance from each satellite in view.
   To calculate its distance from a satellite, a GPS device applies this formula to the satellite's signal: distance = rate x time, where rate is (c) and time is how long the signal traveled through space.
   The signal's travel time is the difference between the time broadcast by the satellite (t1) and the time the signal is received (t2).
iv. Once a GPS device knows its distance from at least four satellites, it can use geometry to determine its location on Earth in three dimensions.

Some GPS applications have the option to set speed limit warning. Generally, the longer the journey and the more measurements that are taken result in a more accurate, overall speed reading.

While some may argue that the speed returned by a GPS receiver is higher than the speedometer in the vehicle, most modern GPS receivers can collect speedometer readings using the OBDII port. This provides a method of verifying the vehicle's speed.

The GPS receiver, fitted with a cellular modem, transmits data to Speed limiter mother board by internal connection normally every minute however it can be more frequent if required. The data is processed on the Speed limiter motherboard.

4. Working of the Over Speed Detecting Device

The system works to determine if a vehicle is exceeding the speed limit based solely on the vehicle's speed? That's where the magic of big data comes in.

The GPS sends the speed calculated based on the aforesaid formula and this speed is send to a program inside the mother board.

The mother board actually works in two parts one a web server and second a continuously running program to monitor the incoming data from the GPS.

The Web Server contains a self-hosted web page which has a simple user interface for input. The input are

i. Mobile number for which the high speed alter should be sent.
ii. Speed Limit in km per hour (kmph) as integer input this will be the check value for the speed limiter.
iii. A save button will be displayed

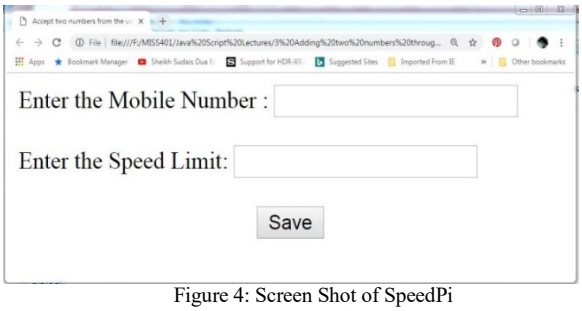


Figure 4: Screen Shot of SpeedPi
Source: Author

The monitor program performs seven actions,

i. Watch for the Speed data from the GPS.
ii. Based on circumference of the wheel calculate the Revolutions per second and convert it to speed. This data can be get from the OBD port if available.
iii. Cross validate the GPS speed data and calculated data.
iv. Check the GPS speed data against the speed limit value entered in the web application.
v. If GPS speed limit crosses the speed limit data then, the program activates another small routine / method which triggers the cellular modem board present alongside the speed limiter motherboard and sends out an SMS to the mobile number entered and stored via web application.
vi. The application continuously send SMS till the speed limit gets below the speed limit mentioned via application.
vii. Once the speed goes below speed limit then the application stops sending SMS and once again starts monitoring the GPS data.

The self-hosted web application can be accessed from any mobile or PC browser.

The speed limiter itself works as a Wi-Fi access point. The mobile or PC can be connected to the speed limiter Wi-Fi network. Then open any browser and type the provided URI in instruction set above. You can see the web application, enter the relevant data and press save.

Then the application captures the data and saves it in a file inside the speed limiter mother board.

The monitor application uses the data stored as file to validate the speed limit as well as mobile number to send SMS.

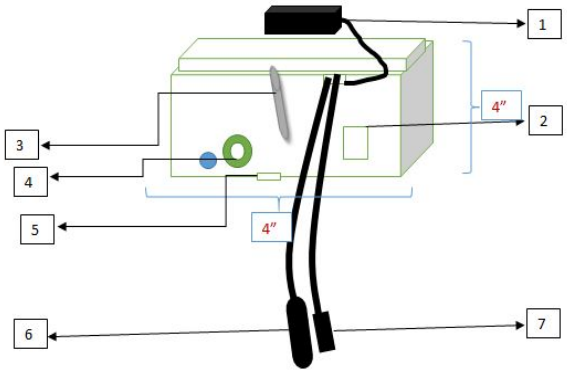


Figure 5: Device Block Diagram
Source: Author

5. Installation and Implementation

1 - GPS receiver
2 - Space for light indication
3 - Mobile Unit Antenna
4 - Mobile unit hard reset button
5 - SIM slot
6 - 12 volts min 2 amp DC power input slot
7 - 05 volts 2 amp DC power input slot

The implementation steps used in detecting the over speed of smart vehicle with the use of IoT is described below with the results:

Step 1: The first step is to insert a valid SIM card with optimal cash balance of SMS in the provided slot.

Step 2: The GPS antenna is placed (a black square concealed piece) in the automobile dashboard or stick it to bonnet (hoop) outside or top of the automobile, it can be affixed to metal because of internal magnet.

Step 3: Power up the Processing Unit and GPS unit with respective 5 volts 2 amperes and 12 volts 2 amperes from the automobile charging unit or utilizing battery with enough amperage.

Step 4: Make a long press for 2 to 3 seconds after powering up the Green button. Item no 4 in reference diagram. This is too difficult and reset the mobile unit.

Step 5: Wait for 4 to 5 seconds to acquire the mobile phone tower and satellite for fixing.

Step 6: A blue button will flash continuously for every two seconds, when viewed and the blue light and one red light can be seen blinking continuously. This mean everything is better to go.

Step 7: Now on the wireless connection a mobile phone is taken and browse for mobile application named SpeedPi, choose it and enter the Raspberry password.

Step 8: Now open the URL in the cell phone and type http://192.168.0.10:8080/Index where a web page can be viewed and the relevant details are entered into it. Care must be taken that they must enter only integers in alert cell number and speed limit any space or characters which will hinder application from performing properly. Then the details are entered and press save a message named OK will be shown.

Step 9: Start the vehicle and enjoy the ride, whenever the limit of speed is crossed an SMS will be send to the cell number entered through the mobile application i.e. SpeedPi

6. Troubleshooting and Maintenance

i. Check the power connections for appropriate amperes using a multi meter.
ii. After finishing step 5 see inside through the spaced provided.

No red light	Check the power
Blinking blue light with blinking green light and static red light	Power off and on the instrument and long press the green button for two to three seconds
No blinking blue	Satellite is not fixed place the instrument in open air
Two blinking blue lights one static red light	Normal works perfect

iii. SMS is not going
   a. Check SIM is correctly inserted in the slot
   b. Check for SMS balance
   c. Check for validity of SIM card
   d. Redo step two (ii) of troubleshooting.
   e. Check for the proper number entered through application
   f. Check for the valid speed limit entered through application
iv. Application is not opening in mobile.
   a. Check for Wi-Fi connection restart the instrument and wait for few seconds and check the availability of Access point **SpeedPi**
   b. Remember to enter proper password **Raspberry**
   c. Ping IP 192.168.0.10 and check for reply.

7. Conclusion

Road accident is a major cause of accidental deaths. Among all the causes of road accidents, over speeding is leading the trail. Several efforts have been made to provide well organized and consistent ways of check and safely handle over speeding. There are many existing approaches but at times they are ineffective. The strategy for this research is to produce a much easily operable and reliable device. Since Over Speeding Detector device is based on radio frequency, the data related to the vehicle is obtained and hence tracking will become easier. Thus it can be concluded that Internet of Things is used in as one of the technology for decreasing the overspeed of vehicle and this technique can be established on roads in future which can be easily transferable.

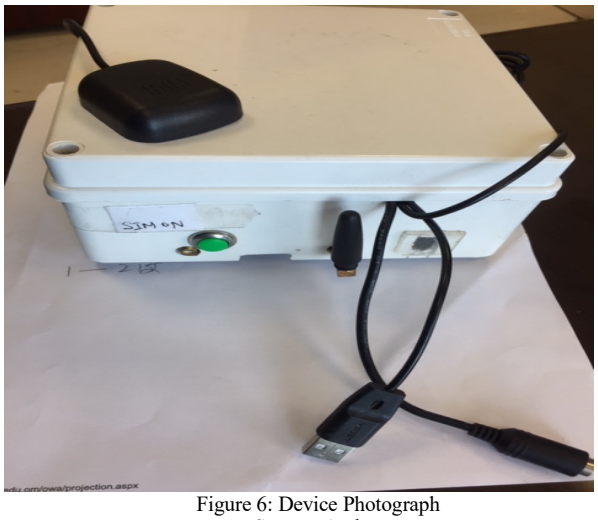


Figure 6: Device Photograph
Source: Author

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was funded and supported by The Research Council (TRC), Sultanate of Oman, we express our thanks and regards to them along with our serving university, Dhofar University, Salalah, Sultanate of Oman. We will also like to express our very sincere thanks to the two reviewers "anonymous", for their insights.

References

[1] Jain, M., Kumar, P., Singh, P., Arora, C. N., & Sharma, A. (2015). Detection of Over Speeding Vehicles on Highways. International Journal of Computer Science and Mobile Computing, 4(4), 613-619.

[2] Deshpande, S., Bhole, V., Dudhade, P., Gourkar, N., & Darade, S. (2017). Implementing a system to detect over speeding & inform authorities in case of any violations, International Research Journal of Engineering and Technology (IRJET), Volume 04, Issue: 04, pp 2445-2449.

[3] Goenka V (2015), How Information Technology can make Indian Roads & Highways Better and Safer!, Available at https://vinitgoenka.wordpress.com/2015/09/25/how-information-technology-can-make-indian-roads-highways-better-and-safer/, accessed on 27th October 2018.

[4] Scroll (2018), Three killed every 10 minutes: Road accident deaths in India up 9% in 4 years, Available at https://scroll.in/article/826264/three-killed-every-10-minutes-road-accident-deaths-in-india-up-9-in-4-years, accessed on 27th October 2018.

[5] Malik Y S (2017), Road Accidents in India-2016, Available at http://www.indiaenvironmentportal.org.in/files/file/Road%20accidents%20in%20India%202016.pdf, accessed on 27th October 2018.

[6] Thakur T T, Naik A, Vatari S and Gogate M (2016), "Real Time Traffic Management using Internet of Things" International Conference on Communication and Signal Processing, pp.6-8.

[7] Sheela. S, Shivaram. K.R, Sunil Gowda.R , Shrinidhi.L, Sahana.S and Pavithra H S (2016), "Innovative Technology for Smart Roads by Using IOT Devices" International Journal of Innovative Research in Science, Engineering and Technology,pp.1-4 Vol. 5, Special Issue 10.

[8] Chandana, K. K, Sundaram M, D'sa C, Swamy M N and Navya K (2017), A Smart Traffic Management System for Congestion Control and Warnings Using Internet of Things (IoT), Saudi Journal of Engineering and Technology, UAE,

[9] Dandala, T. T., Krishnamurthy, V., & Alwan, R. (2017), Internet of Vehicles (IoV) for traffic management. In Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on (pp. 1-4). IEEE.

[10] Barth, M., Boriboonsomsin, K., & Wu, G. (2013), The potential role of vehicle automation in reducing traffic-related energy and emissions. In 2nd IEEE International Conference on Connected Vehicles and Expo (ICCVE),pp. 604-605.

[11] Thakur T and Gogate M (2015), Traffic Controlling and Monitoring using Internet of Things, International Journal of Latest Trends in Engineering and Technology, Volume 6, Issue 2, pp 240-246.

[12] Bharath Kumar Perumalla and M. Sunil Babu (2015), An Intelligent Traffic and Vehcile Monitring System using Internet of Things Architecture, International Journal of Science and Research, Volume 5, Issue 11, pp 853-856

[13] Lakshminarasimhan M (2016), IoT Based Traffic Management System, Available at https://www.researchgate.net/publication/310036684_IoT_Based_Traffic_Management_System, accessed on 26th October 2018.

[14] Sonali P. Kshisagar, Priyanka H. Mantala, Gayatri D. Parjane and Kalyani G. Teke, Intelligent Traffic Management based on IoT, International Journal if Computer Application, Volume 157, Issue 2, pp 26-28

[15] Vijetha H and Nataraj K R (2017), IOT Based Intelligent Traffic Control System, International Journal for Research in Applied Science and Engineering Technology, Volume 5, Issue 5, pp 707-710.

[16] Rath M (2018), Smart Traffic Management System for Traffic Control using Automated Mechanical and Electronic Devices, IOP Publishing Private Limited.

[17] Sharma S, Giradkar V, Sanap A and Sarolkar S (2018), IOT based Traffic Light Controller in Smart City, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Volume 7, Issue 1, pp 117-123.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 1, 171-182 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A Novel Pulse Position Modulator for Compressive Data Acquisition

Constantine A. Pappas*

ECE, Stevens Institute of Technology, 07030, USA

ARTICLEINFO

Article history:
Received:05 December, 2018
Accepted: 28 January, 2019
Online: 11 February, 2019

Keywords:
Nonuniform Sampling
Compressed Sensing
Data Converter

ABSTRACT

This work extends the development of the nonuniform Parallel Digital Ramp Pulse Position Modulation Analog-to-Digital Converter (PDR-ADC) architecture. The continuous to discrete transform of the PDR-ADC is achieved by partitioning the signal amplitude axis into P nonoverlapping partitions that sample the analog input at input signal driven instances. Each partition contains L uniform levels with different quantization step sizes such that the dynamic range of the partitions are related as a geometric series. It is shown that this new architecture satisfies the Nyquist requirement on average (Beutler's condition) and results in a random additive sampling architecture that is alias free (Shapiro-Silverman condition). Additionally, it is shown that the geometric partitioning causes the signal-to-quantization noise ratio (SQNR) to remain approximately constant. A comprehensive design paradigm is presented, including circuits to affect the desired response, the format of the encoded digital samples and the corresponding transformation to determine the equivalent analog voltage. Lastly, although the thrust of this paper is not reconstruction techniques, reconstruction is, nevertheless, compulsory, and recovery and reconstruction is demonstrated through simulations.

1 Introduction

This communication is an extension of work originally presented at a conference on Electrical and Computer Engineering [1] and this current paper significantly expands upon the initial concept proposed in the original material. Although the most common form of sampling is uniform sampling, there are many cases where nonuniform sampling arises and is intentional [2]. Compressed Sensing (CS) ,[3, 4], is based upon, and operates on nonuniform samples. In CS, the goal is to compress at the time of sampling [5]. To combine acquisition and compression into one step necessitates new hardware innovations. The data converter proposed is a novel approach to nonuniform data acquisition.

The mathematical theory of nonuniform sampling and reconstruction has been well studied [6], and several hardware realizations have been described. In [7], the concept of the Level Crossing (LC) detector for nonuniform sampling was described and developed, in [8], the LC concept was extended and [9], an LC hardware design in $120nm$ CMOS was fabricated. In [10], the LC concept was extended to include a triangular dither signal on the input. In [11], the LC sampling scheme was used in an event driven ADC application for electrocardiogram (ECG) signal acquisition. In [12], an adaptive LC sampling scheme was developed, whereby the levels are no longer static but rather adapt to the required signal dynamic range. In [13], a time based ADC was proposed using pulse position modulation (PPM). In [14], a nonuniform sampling system based upon PPM using a reference ramp rest was proposed. In [15], a wideband nonuniform sampling system using a random modulator pre-integrator, similar to direct sequence spread spectrum, was described, and in [16] a nonuniform sampling system based upon pseudo-randomly (PN) triggering the sample-and-hold circuit in an otherwise standard ADC was proposed.

The new nonuniform architecture presented here is based on a parallel implementation of the standard Digital Ramp Pulse Position Modulators (PDR-ADC). The architecture partitions the amplitude range into $\mathcal{P}$ nonoverlapping partitions, each governed by its own digital ramp. The ensemble of digital ramps operate from a single counter N bit counter and a single Digital-

*Corresponding Author: Constantine Pappas, cpappas@stevens.edu

https://dx.doi.org/10.25046/aj040117

to-Analog (DAC) converter, greatly simplifying the synchronization and calibration process. It is shown that our approach exhibits effective compressed sensing performance (compression and acquisition at the same time) at greatly reduced complexity.

2 Conventional PPM

Conventional Pulse Position Modulation (PPM) generates one sample per period of a reference ramp [1]. In the PPM waveform generator circuit shown in Figure 1, $f(t)$ is the input analog signal to be sampled and $r(t)$ is a saw tooth reference waveform with period, T_{ramp}. The comparator is assumed to be referenced from a positive voltage supply, V_{CC}, and a negative supply voltage supply V_{EE}. Pulse Width Modulation (PWM) is an ouput signal and PPM is the output Pulse Position Modulation signal generated by the monostable multivibrator (one shot) circuit. When triggered, a one shot produces a single pulse of fixed, finite duration.

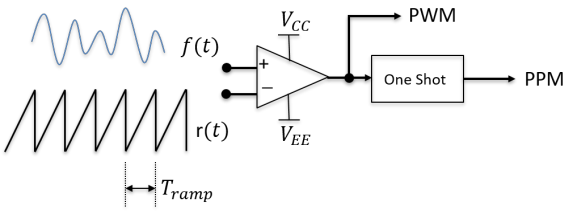


Figure 1: PPM Generator

A comparator, in general, will produce an output at the positive supply rail, V_{CC}, whenever the signal input to the noninverting amplifier terminal, $V_{\oplus}$, is greater than the signal input to the inverting amplifier terminal, $V_{\ominus}$. Similarly, the comparator output will be at the negative supply rail, V_{EE}, whenever $V_{\ominus} > V_{\oplus}$. For the circuit in Figure 1, the PWM output signal as a function of time may be expressed as:

$$PWM(t) = \begin{cases} V_{CC} & \text{if } f(t) > r(t) \\ V_{EE} & \text{otherwise} \end{cases} \tag{1}$$

Under the conditions specified, when $r(t)$ exceeds $f(t)$, the one shot will trigger and a sample generated. Assuming the bandwidth of the analog input, $f(t)$, is less than $\frac{1}{T_{ramp}}$, one sample is generated per period of the reference ramp and we may define the sample rate as, $FS = \frac{1}{T_{ramp}} \equiv \frac{1}{T_s}$.

Figure 2 is an illustration of several cycles of the reference ramp signal. In Figure 2, r_n is the n^{th} period of the reference signal. If we define the full scale voltage as, $V_{FS} = V_{CC} - V_{EE}$ and let the negative supply rail be such that, $V_{EE} = -V_{CC}$, as is typically the case, then we may write, $V_{FS} = 2V_{CC}$. We can then define the slope of the reference as, $\beta = \frac{V_{FS}}{T_s}$. By direct enumeration, $r_n(t)$ in Figure 2 is:

$$\begin{aligned} r_0(t) &= \beta t + V_{EE} && 0 \le t < t_1 \\ r_1(t) &= \beta t + V_{EE} - \beta t_1 && t_1 \le t < t_2 \\ r_2(t) &= \beta t + V_{EE} - \beta t_2 && t_2 \le t < t_3 \\ &\vdots \\ r_n(t) &= \beta t + V_{EE} - \beta t_n && nT_s \le t < (n+1)T_s \end{aligned} \tag{2}$$

Substituting $V_{EE} = -\frac{V_{FS}}{2}$, $V_{FS} = \beta T_s$ and $t_n = nT_s$, (2) becomes:

$$\begin{aligned} r_n(t) &= \beta t - \frac{\beta T_s}{2} - \beta n T_s \\ &= \beta t - \beta T_s\left(n + \frac{1}{2}\right) \qquad n = 0, 1, 2, \cdots \end{aligned} \tag{3}$$

The comparator triggers when $r(t) = f(t)$, from (3), the discrete sample time, τ_n, of the n^{th} pulse is:

$$\tau_n = nT_s + \frac{f_n}{\beta} + t_d \qquad n = 0, 1, 2, \cdots \tag{4}$$

where $\beta = \frac{V_{FS}}{T_s}$ and $t_d = \frac{T_s}{2}$.

Equation (4) shows, in conventional PPM not only is the timing of the n^{th} sample proportional to the sample number, n, the timing is also a function of the signal being sampled. Consequently, it is seen, conventional PPM generates nonuniform sampling.

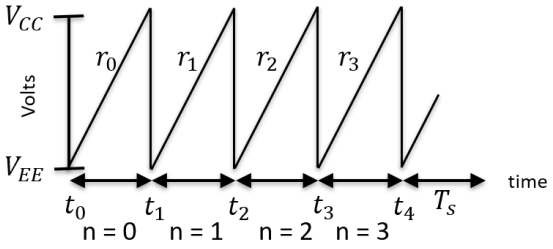


Figure 2: PPM Reference Ramp

3 Δ-PPM

It is possible to generate a PPM signal directly, without explicitly having the PWM output trigger a one-shot. Such a direct generation of the PPM signal is accomplished by resetting the reference ramp after the comparator triggers. Generating a PPM signal by resetting the reference ramp will be called, Δ–PPM, to distinguish it from conventional PPM. Δ–PPM may be generated with the circuit of Figure 3. In Figure 3, V_{CC} and V_{EE} are the power supplies for the comparator, $f(t)$ is the signal to be sampled and $r(t)$ is the reference ramp. Also shown in Figure 3 is a RESET circuit that asynchronously resets the reference ramp at each PPM pulse.

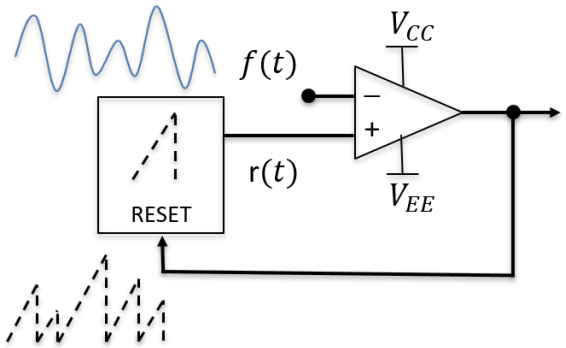


Figure 3: Reset PPM Generator

The previous periodicity of the reference signal is annihilated when the reference ramp is allowed to reset after the comparator triggers. An arbitrary response to the asynchronous triggering is shown in Figure 4. In Figure 4, we call r_n is the n^{th} response of the reference signal, rather than the n^{th} period of the reference, and again define the full scale voltage as, $V_{FS} = V_{CC} - V_{EE}$ and the slope of the reference as, $\beta = \dfrac{V_{FS}}{T_s}$.

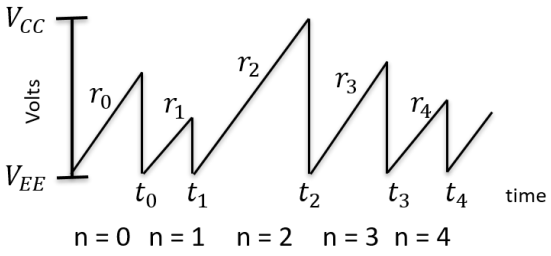


Figure 4: Reset PPM Reference Ramp

By direct enumeration, $r_n(t)$ in Figure 4 is:

$$\begin{aligned} r_0(t) &= \beta t + V_{EE} & 0 \le t < t_0 \\ r_1(t) &= \beta t + V_{EE} - \beta t_0 & t_0 \le t < t_1 \\ r_2(t) &= \beta t + V_{EE} - \beta t_1 & t_1 \le t < t_2 \\ &\vdots \\ r_n(t) &= \beta t + V_{EE} - \beta t_{n-1} & t_{n-1} \le t < t_n \end{aligned} \tag{5}$$

We note that the interval endpoints cannot be specified as constants, as was the case in conventional PPM, because they evolve dynamically. Again, substituting $V_{EE} = -\dfrac{V_{FS}}{2}$ with $V_{FS} = \beta T_s$, (5) becomes:

$$r_n(t) = \beta t - \beta t_{n-1} - \frac{\beta T_s}{2} \qquad t_{n-1} \le t < t_n \tag{6}$$

The comparator triggers when $r(t) = f(t)$, from (6), the discrete sample time, τ_n, of the n^{th} pulse is:

$$\tau_n = \tau_{n-1} + \frac{f_n}{\beta} + t_d \qquad n = 1, 2, 3, \cdots \tag{7}$$

where $\beta = \dfrac{V_{FS}}{T_s}$ and $t_d = \dfrac{T_s}{2}$.

Equation (7) shows, in Δ–PPM the timing of the n^{th} sample is a function of the signal being sampled and thus Δ–PPM generates nonuniform sampling, as in conventional PPM. Additionally, from (7), Δ–PPM samples times are a function of the previous sample time.

Two important consequence result if the sample times are a function of the previous sample time. First, Shapiro and Silverman [20] showed that sampling can be made alias free if each sample time is derived from the previous one by the addition of an independent random variable, from (7) it is seen that Δ–PPM provides such a sampling scheme. Second, Beutler [21] showed that if the sampling rate exceeds the Nyquist rate *on average* then the nonuniform sampling set will be stable and can be used to reconstruct a band-limited signal. We now show that Δ–PPM produces a sampling set with an average sampling rate,R_{avg}, that approaches, $R_{avg} = \dfrac{2}{T_s} \approx 2FS$.

Using (7) iteratively we may write:

$$\tau_n = \tau_o + n\frac{T_s}{2} + \frac{1}{\beta}\sum_{k=1}^{k=n} f_k \tag{8}$$

where τ_o is the first sampling instant.

Let $n \to N$, be the total number of samples such that $\{N : N \in 1, 2, 3, \ldots \infty\}$ and let τ_N be the maximum sample time, then, dividing by N:

$$\frac{\tau_N}{N} = \frac{\tau_o}{N} + \frac{T_s}{2} + \frac{1}{\beta}\left(\frac{1}{N}\sum_{k=1}^{k=N} f_k\right) \tag{9}$$

The last term in (9) is the average value of $f(t)$. If the average value equals zero, then the average sample rate is given by:

$$R_{avg} = \frac{2}{T_s}\left(\frac{N}{N + 2\tau_o/T_s}\right) \tag{10}$$

The maximum value of the first sampling instant is, $\tau_o = T_s$, which corresponds to the minimum value of R_{avg}, thus:

$$R_{avg} \ge \frac{2}{T_s}\left(\frac{N}{N+2}\right) \tag{11}$$

Less formally, for reasonable values of N experienced in practice, we may regard the average sample rate in Δ-PPM as, $R_{avg} = \dfrac{2}{T_s} \approx 2FS$.

The significance of (7) and (11) are that Δ–PPM is *self-regulating*. Δ–PPM automatically satisfies the Nyquist requirement on average (Beutler's condition [21]) and produces alias free random sampling (Shapiro-Silverman condition [20]). Additionally, Δ–PPM achieves self-regulation with no a priori knowledge of the signal support and does not utilize any particular code or special sequence to generate random samples.

4 Geometric Partitioning

The stylized, equal partitioning, presented [1], was intended to provide an introduction for the essence of the PDR-ADC. We now develop the elaboration of the partitioning scheme, where significant benefits will be obtained.

All data converters need an analog reference voltage to accomplish the continuous to discrete transform and produce a digital word [22]. In general, the converter divides an analog reference voltage, V_{REF}, into a fixed number of analog voltage levels, L. These analog voltage levels are then mapped to a digital number, typically referred to as counts. By convention, the smallest analog voltage level is assigned digital level $\boxed{0}$. The remaining analog voltage levels are mapped to digital levels by incrementing the ADC count. In this way, it is not possible to map the analog reference voltage, V_{REF}, to a number that can be reached and assigned a digital count.

All of the digital levels taken together define the scale. In general, there will be L levels and $L-1$ steps. The reference voltage divided by the number of levels defines the size of the analog step to reach an adjacent level. The step size, Δ, is sometimes referred to as the quantization step size or the least significant bit (LSB). The maximum analog voltage that can be mapped to a digital level is called the *full scale* voltage, V_{FS}. The relationship between the reference voltage, the full scale voltage and the quantization step size are shown in Figure 5 for a converter with $L = 16$ levels.

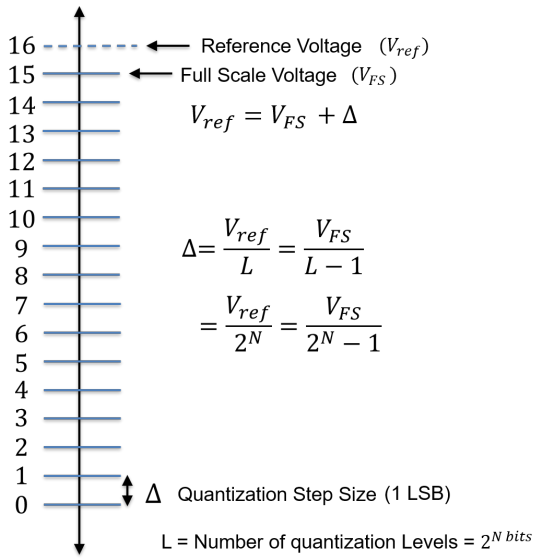


Figure 5: Quantization Step Size

The number of levels, L, is typically designed to be a function of the number of bits as, $L = 2^N$, where N is the number of bits. For the 4 bit converter shown in Figure 5, the reference voltage is divided by $16 = 2^4$. The first digital level is assigned digital count $\boxed{0}$, and corresponds to the digital number, $\boxed{0}\boxed{0}\boxed{0}\boxed{0}$. The maximum digital count is $\boxed{15}$, and corresponds to the digital number, $\boxed{1}\boxed{1}\boxed{1}\boxed{1}$. It is not possible to encode digital count $\boxed{16}$ with a 4 bit counter, and consequently, the reference voltage, V_{REF} in Figure 5, is not mapped to a digital number.

The principle of the geometric partitioning is shown in Figure 6 for a system with $\mathcal{P} = 8$ partitions and $L = 4$ levels. In the PDR data converter, the signal amplitude axis in partitioned into $\mathcal{P}$ partitions such that each partition contains L levels, the partitions do not overlap and each partition has a different quantization step size. The geometric partitioning is achieved by relating the spans (the dynamic range) of the partitions as a geometric series.

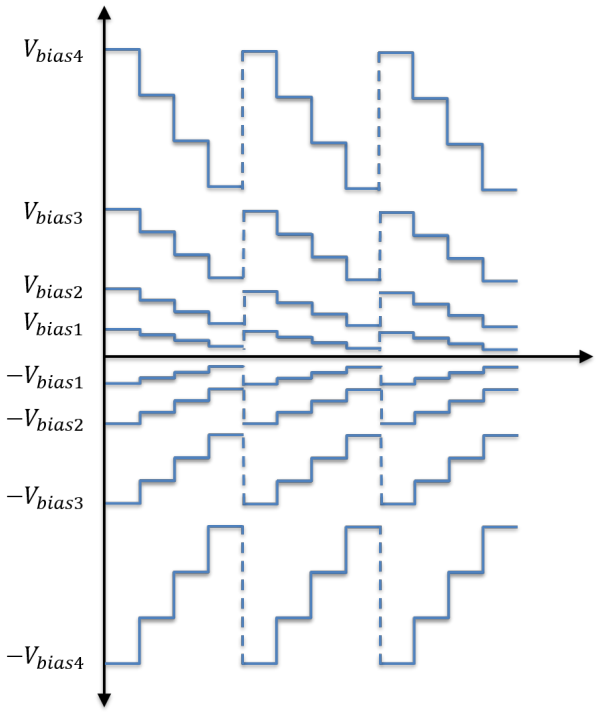


Figure 6: Geometric Partitioning: Conceptual

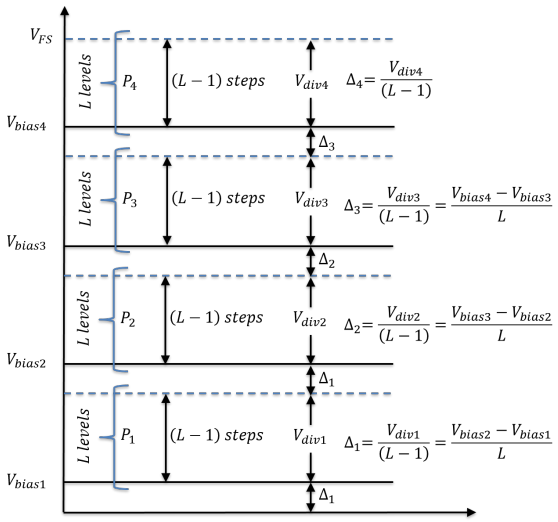


Figure 7: Geometric Partitioning: Detail

The relationship between partition bias voltages, V_{bias_m}, and the quantization step sizes needed to realize the behavior shown in Figure 6, can be better understood and visualized with the aid of Figure 7. Due to the symmetry of the bias voltages, only the positive partitions, P_1, P_2, P_3 and P_4, as shown in Figure 7, are needed.

Let the total number of partitions, $\mathcal{P}$, be even, and define the maximum partition number to be, $\mathcal{M} = \frac{\mathcal{P}}{2}$, and let m denote the m^{th} partition. In Figure 7, we denote by V_{div_m} the span of the m^{th} partition. By design, the spans of the partitions are geometrically related, thus:

$$\begin{aligned} V_{div2} &= 2V_{div1} = 2^{(2-1)}V_{div1} \\ V_{div3} &= 4V_{div1} = 2^{(3-1)}V_{div1} \\ V_{div4} &= 8V_{div1} = 2^{(4-1)}V_{div1} \\ &\vdots \\ V_{div_m} &= 2^{(m-1)}V_{div1} \quad m = 1, 2, 3, \cdots \mathcal{M} \end{aligned} \tag{12}$$

To realize (12), a circuit that takes the output from the DAC and applies the appropriate gain to compress the DAC steps is required. The required step compression response is shown in Figure 8 and the circuit to realize the response is shown in Figure 9.

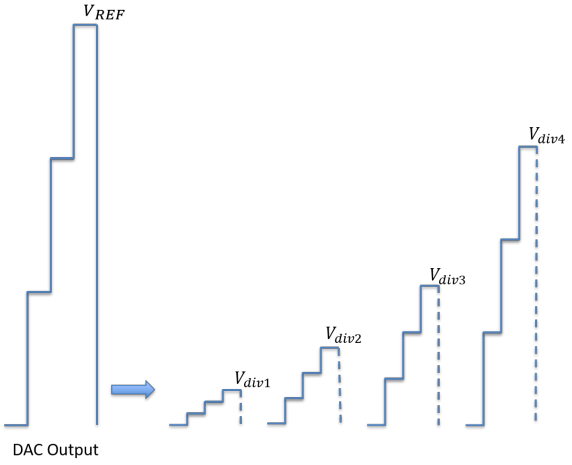


Figure 8: Step Compression Voltages

The step compression circuit in Figure 9 is a non-inverting voltage divider. To obtain design equations for the bias voltages in terms of the reference voltage, V_{REF}, the maximum number of partitions, $\mathcal{M}$ and the number of levels, L, we design the step compression circuit such that the voltage drop across resistor, $V_{Rx} = \left(2^{(\mathcal{M}-1)} - 1\right)V_{div1}$. Then, due to the geometric design and Kirchoff's voltage law, it must be the case:

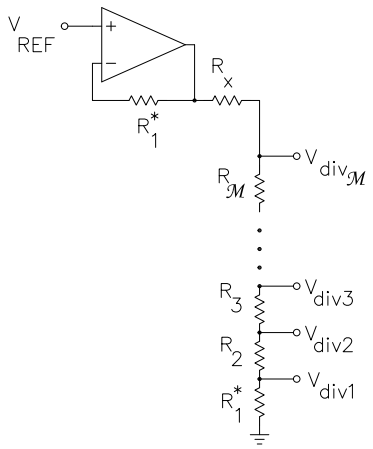


Figure 9: Step Compression Circuit

$$\begin{aligned} V_{REF} &= V_{div1} + V_{div2} + V_{div3} + \cdots + V_{div\mathcal{M}} \\ V_{REF} &= V_{div1}\left(1 + 2 + 4 + 8 + \cdots + 2^{(\mathcal{M}-1)}\right) \\ V_{REF} &= V_{div1}\sum_{k=0}^{k=\mathcal{M}-1} 2^k \\ V_{REF} &= V_{div1}\left(2^{\mathcal{M}} - 1\right) \\ &\therefore \\ V_{div1} &= \frac{V_{REF}}{2^{\mathcal{M}} - 1} \end{aligned} \tag{13}$$

From (12) and (13), the m^{th} voltage divider voltage is given by:

$$V_{div_m} = \left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1}\right)V_{REF} \tag{14}$$

Dividing (14) by $L-1$, the quantization step size, Δ, is:

$$\Delta_m = \left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1}\right)\left(\frac{V_{REF}}{L-1}\right) \tag{15}$$

Using Figure 7 and (14) and (15), we may write the bias voltages, V_{bias_m}, as:

$$\begin{aligned} V_{bias_m} &= \Delta_1 + \sum_{k=1}^{k=m-1}\left(V_{div_m} + \Delta_m\right) \\ V_{bias_m} &= \Delta_1 + \\ &\quad + \sum_{k=1}^{k=m-1}\left(\frac{2^{(m-1)}}{2^{\mathcal{M}} - 1}V_{REF} + \frac{2^{(m-1)}}{2^{\mathcal{M}} - 1}\frac{V_{REF}}{L-1}\right) \\ V_{bias_m} &= \Delta_1 + \frac{V_{REF}}{2^{\mathcal{M}} - 1}\left(\frac{L}{L-1}\right)\sum_{k=1}^{k=m-1} 2^{k-1} \\ V_{bias_m} &= \Delta_1 + \frac{V_{REF}}{2^{\mathcal{M}} - 1}\left(\frac{L}{L-1}\right)\left[2^{(m-1)} - 1\right] \\ V_{bias_m} &= \frac{V_{REF}}{2^{\mathcal{M}} - 1}\left(\frac{L}{L-1}\right)\left(2^{(m-1)} - 1 + \frac{1}{L}\right) \end{aligned} \tag{16}$$

Lastly, the Full Scale voltage shown in Figure 7 may be obtained by adding the results of (16) and (14) evaluating at $m = \mathcal{M}$:

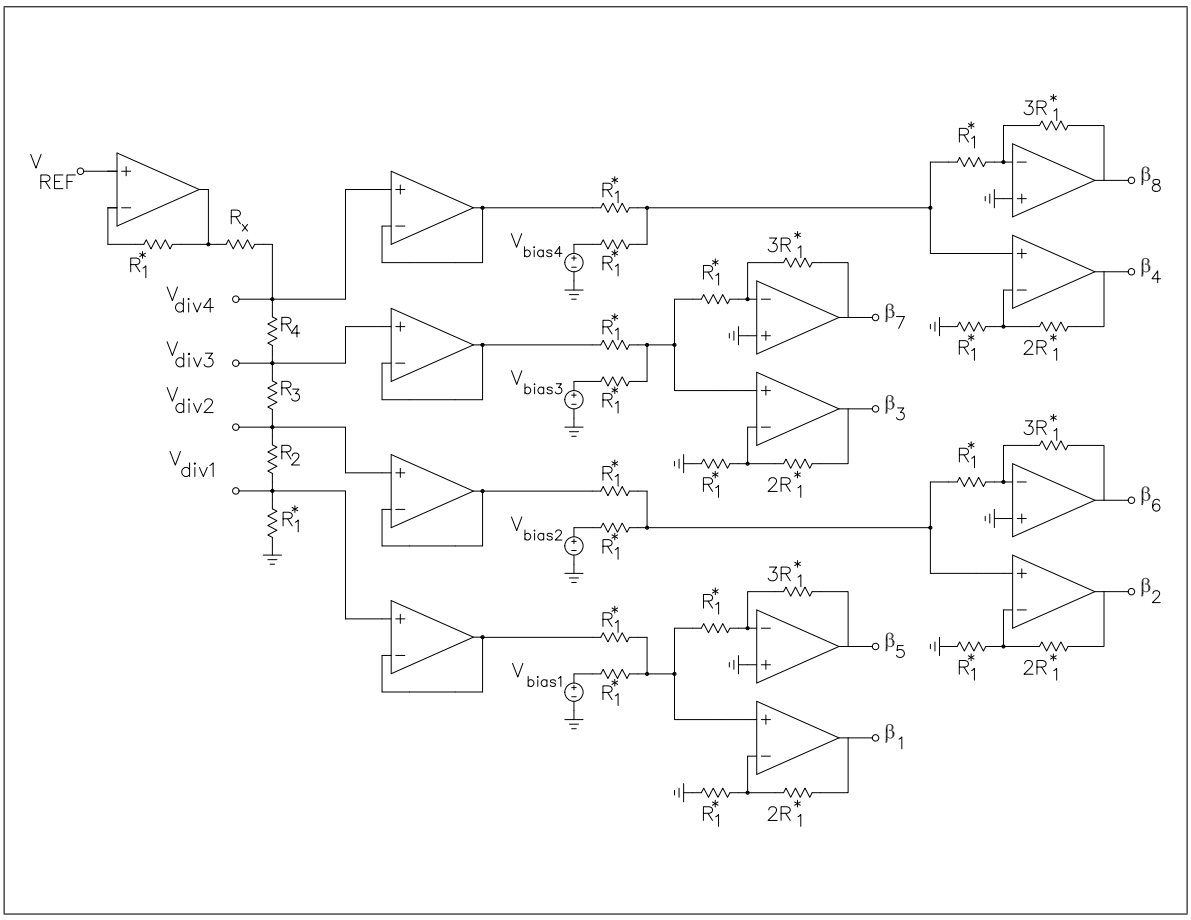


Figure 10: Step Compression & Level Shifting Circuit

$$V_{FS} = \frac{V_{REF}}{2^{\mathcal{M}}-1}\left(\frac{L}{L-1}\right)\left[2^{\mathcal{M}}-1+\frac{1-2^{(\mathcal{M}-1)}}{L}\right] \tag{17}$$

In terms of the external design parameters, V_{REF}, $\mathcal{M}$ and L, the number of partitions, $\mathcal{M}$ will have the most influence on the design equations, (16) and (17), since the bias is proportional to $\dfrac{1}{2^{\mathcal{M}}-1}$.

5 Step Compression and Level Shifting

We now develop the circuit to generate the parallel digital ramps shown in Figure 6, and we call this function, step compression and level shifting (SCLS). The step compression and level shifting circuit is shown in Figure 10. Step compression is governed by equations (14) and (15) and we seek now to determine design equations for the resistance for the circuit in Figure 10.

5.1 Step Compression

The step compression gains can be realized from circuit analysis by solving for the resistor values, in Figure 8, required to establish the voltage divider voltages, V_{div_m} as given by (14).

Let $\sum R_m = R_1^* + R_2 + R_3 + \cdots + R_{\mathcal{M}}$, then from circuit analysis:

$$V_{div1} = V_{REF}\left(\frac{R_1^*}{R_x + \sum R_m}\right) \tag{18a}$$

and in general:

$$V_{div_m} = V_{REF}\left(\frac{R_1^* + R_2 + R_3 + \cdots + R_m}{R_x + \sum R_m}\right) \tag{18b}$$

Divide (18b) by (18a):

$$\frac{V_{div_m}}{V_{div1}} = \frac{R_1^* + R_2 + R_3 + \cdots + R_m}{R_1^*} \tag{19}$$

Substituting the geometric relation given in (12) for V_{div_m}:

$$2^{(m-1)}R_1^* = R_1^* + R_2 + R_3 + \cdots + R_m = \sum R_m \tag{20a}$$

from which it is seen:

$$\begin{aligned} m = 2 &\longrightarrow R_2 = R_1^* \\ m = 3 &\longrightarrow R_3 = 2R_1^* \\ m = 4 &\longrightarrow R_4 = 4R_1^* \\ &\vdots \end{aligned} \tag{20b}$$

and in general:

$$R_m = 2^{(m-2)}R_1^* \quad \text{for} \quad m > 1 \tag{20c}$$

Lastly, from Equation (13) and (18a):

$$\frac{1}{2^{\mathcal{M}}-1} = \frac{R_1^*}{R_x + \sum R_{\mathcal{M}}} \qquad R_x = \left(2^{\mathcal{M}}-1\right)R_1^* - \sum R_{\mathcal{M}} \tag{21a}$$

and from (20a):

$$R_x = \left(2^{\mathcal{M}}-1\right)R_1^* - 2^{(\mathcal{M}-1)}R_1^* \qquad R_x = \left[2^{(\mathcal{M}-1)}-1\right]R_1^* \tag{21b}$$

To complete the design of the Step Compression circuit, we must specify a value for resistor R_1^*. To do so, we seek a design equation that relates the RMS noise voltage of the resistor to the voltage of the smallest quantization step size, Δ_1 in Figure 7. We then design the resistor value be *less* than this value so that the system is limited by the quantization noise of the LSB rather than the thermal noise of the resistor.

From (15):

$$\Delta_1 = \frac{V_{REF}}{(2^{\mathcal{M}}-1)(L-1)} \tag{22}$$

Assuming the noise is Gaussian distributed, then approximately all of the noise is contained within 6.6 standard deviations[1]. The peak-to-peak thermal noise voltage of the resistor is given by:

$$V_{pp\mathcal{N}_{th}}|_{R_1^*} = 6.6\sqrt{(4RkT\mathcal{B})} \tag{23}$$

where R is the resistance in Ohms, k is Boltzmann's constant ($k \approx 1.30865\times10^{-23}J/K$), T is the temperature in Kelvin and $\mathcal{B}$ is the bandwidth in Hertz.

To design for the thermal noise voltage of R_1^* to be less than Δ_1, we should select R_1^*, such that:

$$R_1^* \le \left(\frac{1}{6.6}\right)^2\left(\frac{V_{REF}}{(L-1)(2^{\mathcal{M}}-1)}\right)^2\left(\frac{1}{4kT\mathcal{B}}\right) \tag{24}$$

5.2 Level Shifting

Level shifting is governed by the bias voltages as given by (16). The aim of the level shifting circuit is to shift the step compressed voltages, shown in Figure 8, to the required bias level, shown in Figure 6. The β_m signals, in Figure 10, are the final step compressed and level shifted signals that are feed back to the input comparators [1]. From the symmetry of the design, the positive shifted β signals are given by:

$$\beta_m = V_{div_m} + V_{bias_m} \quad \text{for} \quad 1 \le m \le \mathcal{M} \tag{25}$$

and the negative shifted signals are given by:

$$\beta_{\mathcal{M}+m} = -\left(V_{div_m} + V_{bias_m}\right) \text{ for } 1 \le m \le \mathcal{M} \tag{26}$$

Specifically, for the β signals shown in Figure 10,:

$$\begin{aligned} \beta_1 &= V_{div1} + V_{bias1} & \beta_5 &= -(V_{div1} + V_{bias1}) \\ \beta_2 &= V_{div2} + V_{bias2} & \beta_6 &= -(V_{div2} + V_{bias2}) \\ \beta_3 &= V_{div3} + V_{bias3} & \beta_7 &= -(V_{div3} + V_{bias3}) \\ \beta_4 &= V_{div4} + V_{bias4} & \beta_8 &= -(V_{div4} + V_{bias4}) \end{aligned} \tag{27}$$

[1] 6σ is often used, however, 6.6σ is becoming an industry standard.

The design is readily obtained if the inverting and noninverting amplifiers in 10 are solved with the feedback resistors as arbitrary unknown resistors (2 degrees of freedom) and the remaining resistors fixed, equal to resistor R_1^*. A pair of inverting and noninverting amplifiers are shown in Figure 11 where $R_{\ominus}$, is the resistor in the inverting amplifier and $R_{\oplus}$ is the resistor in the non-inverting amplifier.

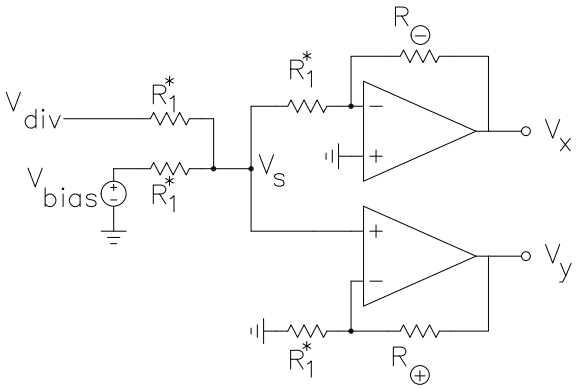


Figure 11: Level Shifting Feedback Scaling

From circuit analysis, the node voltage, V_s is:

$$\frac{V_s - V_{div}}{R_1^*} + \frac{V_s - V_{bias}}{R_1^*} + \frac{V_s}{R_1^*} = 0 \qquad V_s = \frac{V_{div} + V_{bias}}{3} \tag{28}$$

For the inverting amplifier we have, $\frac{V_s}{R_1^*} = -\frac{V_x}{R_{\ominus}}$, from which:

$$V_x = -\frac{R_{\ominus}}{R_1^*}\left(\frac{V_{div} + V_{bias}}{3}\right) \tag{29}$$

The required inverting sum, as given in (27), is obtained when:

$$R_{\ominus} = 3R_1^* \tag{30}$$

Similarly, for the non-inverting amplifier we have: $V_y = \left(1 + \frac{R_{\oplus}}{R_1^*}\right)V_s$, from which:

$$V_y = \left(1 + \frac{R_{\oplus}}{R_1^*}\right)\left(\frac{V_{div} + V_{bias}}{3}\right) \tag{31}$$

The required non-inverting sum, as given in (27), is obtained when:

$$R_{\oplus} = 2R_1^* \tag{32}$$

6 Maximum Sample Rate

Samples are generated by the precision windowed dual slope edge detector shown in Figure 12. The edge detector triggers an asynchronous LOAD flip-flop that stores the instantaneous value of the counter to memory. The LOAD signal, additionally, "sets" an RS flip-flop that drives the synchronous RESET flip-flop. On the next CLOCK, T_{clk}, the RESET flip-flop resets the counter. The counter generates a HOLD OFF signal that inhibits additional LOAD signals until the counter has settled, where the safety time to settle is 1 CLOCK pulse. The sampling circuit is shown in Figure 12.

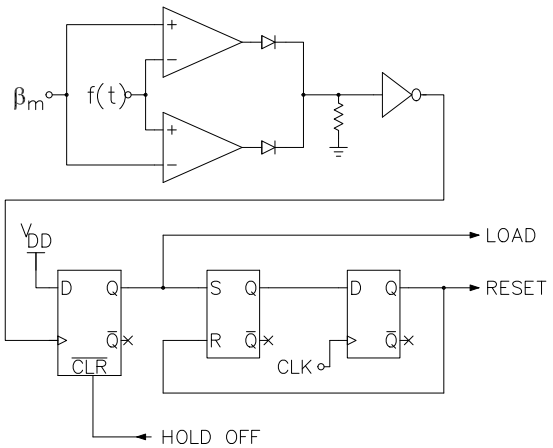


Figure 12: Windowed Synchronous One-Shot

In the PDR data converter, we must ensure that the counter has settled before the next LOAD/RESET cycle. The worst case signaling (fastest signal) corresponds if the LOAD signal aligns with a CLOCK edge. In such a case, a minimum of 3 CLOCKS is required to guarantee correct conversion before the next LOAD signal is allowed to register a new sample, this worst case timing is shown in Figure 13.

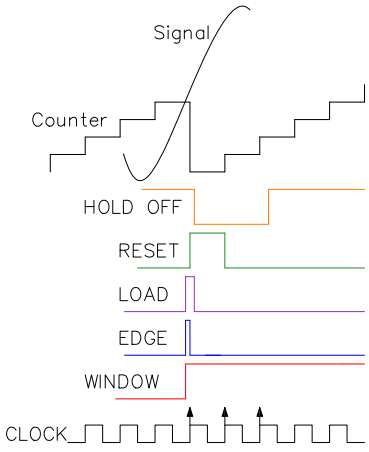


Figure 13: LOAD/RESET Timing

To determine the maximum signal frequency that the PDR-ADC can accommodate, we equate the maximum rate of change of the input signal, to the maximum rate of change allowable by the PDR. With an input sinusoid of the form, $f(t) = V_{FS} sin(2\pi f_{sig} t)$, the maximum rate of change is: $\frac{\Delta V}{\Delta t} = 2\pi f_{sig} V_{FS}$. The fastest rate of change the PDR can respond to is: $\frac{\Delta V}{\Delta t} = \frac{2V_{FS}}{3T_{clk}}$. The maximum signal frequency is:

$$f_{sig_{max}} = \frac{1}{3\pi T_{clk}} \tag{33}$$

7 The Counting Vector

In this section, we establish some important properties of the counting vector. Each intersection of $f(t)$ with a reference counter step contributes to the counting vector, $\boldsymbol{\alpha}$. The counting vector is responsible for determining the time of each sample, the amplitude of each sample, and the number of samples acquired. The counting vector also contains the information about the number of missing samples and where these missing samples are located. Information regarding the missing samples is critically important for reconstruction as the location and number of the missing samples (the data to be interpolated) must be known.

We denote by α, the number of counts accumulated by the counter. If the counter is strictly counting up, then the time to accumulate α counts, $t_\alpha = (\alpha + 1)T_{clk}$, where T_{clk} is the clock period. The time of the n^{th} sample, $t_{[n]}$, is the cumulative sum of the t_α's:

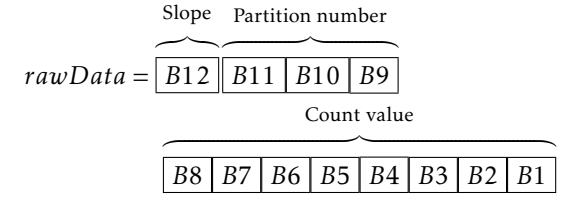
$$t_{[n]} = T_{clk} \sum_{i=1}^{i=n} (\alpha_i + 1) \tag{34}$$

If, instead, the counter is strictly counting down, we denote by α_c (the complement of α) the number of counts accumulated by the counter. In this way, $\alpha_c = (L-1) - \alpha$, where $L-1$ is the maximum value of the counter. In this case, the time of the n^{th} sample is:

$$t_{[n]} = T_{clk} \sum_{i=1}^{i=n} (L - \alpha_i) \tag{35}$$

The direction of the counter can be changed, by using a toggle flip-flop clocked on each RESET signal, and the system will continue to maintain the correct timing of each sample.

To recover the signal requires a method to resolve the counter slope, the partition number and the count value. We append, to the counting vector, additional bits that correspond to the counter slope and the partition number. As a specific example of accommodating partition encoding, suppose a PDR-ADC is designed with $\mathcal{P} = 8$ partitions and $L = 256$ levels per partition. The partition encoder requires 3 bits, the count slope requires 1 bit and the count value requires 8 bits, the data word stored in memory will be of the form:



In this example, the partition number in each data word is determined by:

$$m_i = rawData_i\ [B_{11} : B_9][\ 4\ 2\ 1\ ]^T \tag{36}$$

and each α_i is determined by:

$$\begin{aligned} \alpha_i &= rawData_i\ [B_8 : B_1]\mathbf{W}^T \\ &\equiv \mathbf{B}_i\mathbf{W}^T \end{aligned} \tag{37}$$

where, $\mathbf{W}$ = [128 64 32 16 8 4 2 1].

In general, once the i^{th} partition number, m_i, and the i^{th} count, α_i, have been determined, the sampled voltage value is given by:

$$V_i = \alpha_i \Delta_{m_i} + V_{bias_{mi}} \tag{38}$$

where Δ_m is given by (15) and V_{bias_m} is given by (16).

Lastly, since $\alpha_i + 1$ is the number of clocks to obtain α_i counts, then α_i is equal to the number of missing samples if the signal had been sampled at a rate equal to T_{clk}.

8 Signal to Quantization Noise Ratio (SQNR)

The signal to noise ratio (SNR) is always equal to the ratio of the signal power, P_{sig}, to the noise power, $P_{\mathcal{N}_Q}$. In data converters, the "noise" added by the system due to the act of approximating (truncating) a continuous function to finite precision is quantization noise, and the ratio of interest is the signal to quantization noise, $SQNR = P_{sig}/P_{\mathcal{N}_Q}$. In uniform quantizers, the quantization noise power is well approximated by, [17], [23]:

$$P_{\mathcal{N}_Q} = \frac{\Delta^2}{12} \tag{39}$$

In the PDR data converter, from (15), for any adjacent partitions:

$$\frac{\Delta_m}{\Delta_{m+1}} = \frac{1}{2} \tag{40}$$

from (39) and (40):

$$P_{\mathcal{N}_{Qm}} = \frac{P_{\mathcal{N}_{Qm+1}}}{4} \tag{41}$$

For the PDR, (41) states, the quantization noise power decreases by a factor of 4 when transitioning from a higher partition to a lower level partition. This effect is shown in the bottom panel in Figure 14, where, for comparison, we have also plotted the quantization error of a uniform quantizer in the top panel.

Let a signal of the form, $y = A_o sin(\omega_o t)$, be input to a uniform quantizer with quantization noise power, $P_{\mathcal{N}_{Qu}}$, the SQNR is given by:

$$SQNR_u = \frac{P_{sig}}{P_{\mathcal{N}_{Qu}}} = \frac{A_o^2}{2}\frac{1}{P_{\mathcal{N}_{Qu}}} \tag{42}$$

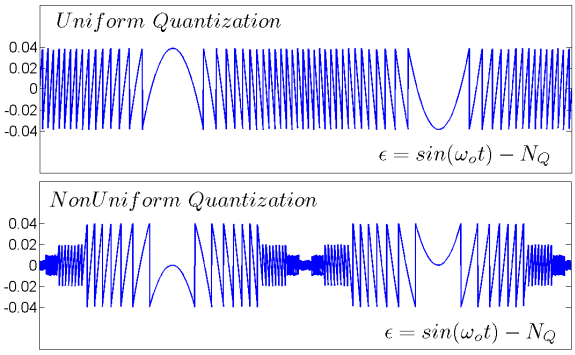


Figure 14: Quantization Error Comparison

Now suppose a signal of the form, $y = \frac{A_o}{2} sin(\omega_o t)$ is input to this uniform quantizer, the SQNR becomes:

$$SQNR_u = \left(\frac{1}{4}\right)\frac{A_o^2}{2}\frac{1}{P_{\mathcal{N}_{Qu}}} \tag{43}$$

From (42) and (43), in a uniform quantizer, when the input signal amplitude decreases by a factor of 2, the SQNR degrades by a factor of 4.

Now consider a signal of the form, $y = A_o sin(\omega_o t)$, input to the nonuniform PDR quantizer with quantization noise power, $P_{\mathcal{N}_{Qnu}}$, the SQNR is given by:

$$SQNR_{nu} = \frac{P_{sig}}{P_{\mathcal{N}_{Qnu}}} = \frac{A_o^2}{2}\frac{1}{P_{\mathcal{N}_{Qnu}}} \tag{44}$$

Again, suppose the input signal amplitude decreases by a factor of 2 and let $y = \frac{A_o}{2} sin(\omega_o t)$ be input to the nonuniform PDR quantizer, then, by (41), the SQNR becomes:

$$\begin{aligned} SQNR_{nu} &= \left(\frac{1}{4}\right)\frac{A_o^2}{2}\frac{1}{P_{\mathcal{N}_{Qnu}}/4} \\ &= \frac{A_o^2}{2}\frac{1}{P_{\mathcal{N}_{Qnu}}} \end{aligned} \tag{45}$$

From (44) and (45), it is seen, the geometric partitioning of the PDR-ADC attempts to maintain the signal-to-quantization noise ratio constant, this is a significant improvement compared to uniform quantization data converters.

We shall now be concerned with determining this constant. In an N bit uniform quantizer, the quantization step size, Δ_u, is:

$$\Delta_u = \frac{2V_{FS}}{2^N - 1} \tag{46}$$

Suppose we have a PDR data converter, with maximum partition number, $\mathcal{M}$, and $L = 2^N$ levels per partition, where N is the same as the uniform quantizer in (46). In the PDR, the largest quantization step size is given by (15), evaluated at $m = \mathcal{M}$.

$$\begin{aligned} \Delta_{\mathcal{M}} &= \left(\frac{2^{(M-1)}}{2^{\mathcal{M}} - 1}\right)\left(\frac{V_{REF}}{2^N - 1}\right) \\ &\approx \left(\frac{1}{2}\right)\left(\frac{V_{REF}}{2^N - 1}\right) \end{aligned} \tag{47}$$

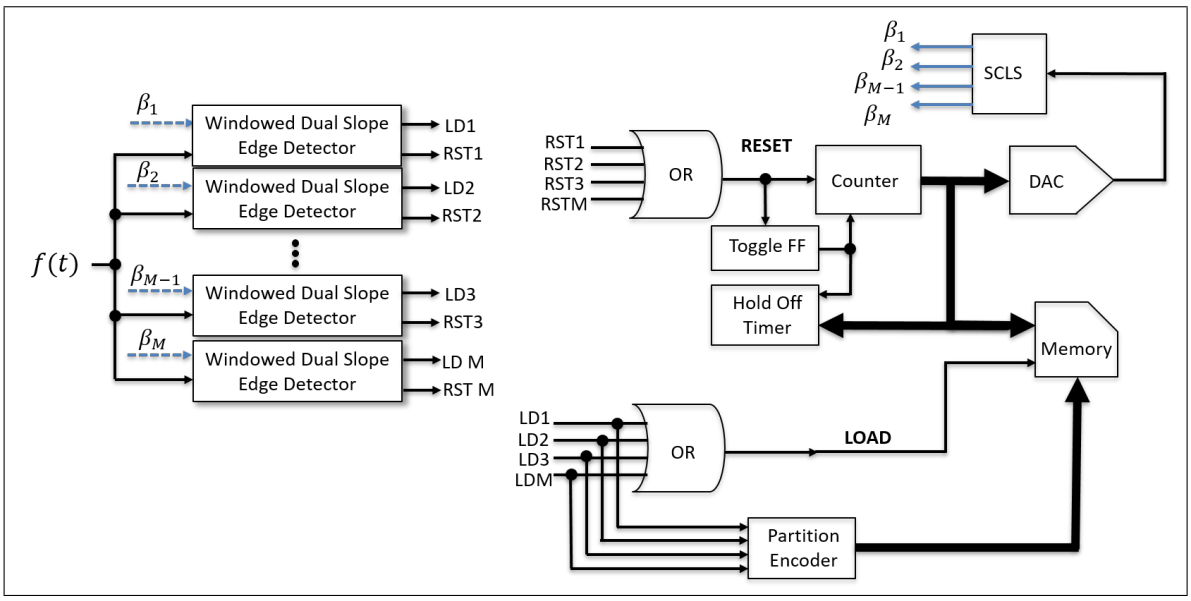


Figure 15: Parallel Digital Ramp ADC Overall Block Diagram

Using (46), we may write (47) as:

$$\Delta_{\mathcal{M}} \approx \left(\frac{1}{4}\right)\left(\frac{2V_{REF}}{2^N - 1}\right) = \frac{2V_{REF}}{2^{(N+2)} - 4} \tag{48}$$

In a PDR data converter, with $L = 2^N$ levels per partition, (48) states, the PDR has gained approximately 2 bits of resolution compared to a uniform quantizer.

The results of (45) and (48) may be summarized as, given a PDR data converter with $L = 2^N$ levels, the SQNR of the system will be approximately equivalent to a uniform system with $L = 2^{(N+2)}$ levels and the PDR will attempt to maintain this performance for all input signal levels.

9 Simulation Results

The proposed parallel digital ramp ADC was modeled and simulated in Simulink® and the reconstruction performed in Matlab®. The overall block diagram of the new ADC is shown Figure 15.

9.1 Linearity

The linearity of the PDR was analyzed empirically using a triangle wave, as shown in Figure 16, since, by design, the sampled data is not equally spaced. This approach was taken because the data converter figure of merit, Differential Non-Linearity (DNL), is a function of the difference of consecutive samples with a constant step size, as given by (49), where for an ideal ADC, the DNL = 0 [18]. In the PDR, since large gaps appear in the data and the step size is not constant, linearity is more easily performed graphically.

$$DNL[k] = \frac{ADC_{out}[k] - ADC_{out}[k-1]}{\Delta_{LSB}} - 1 \tag{49}$$

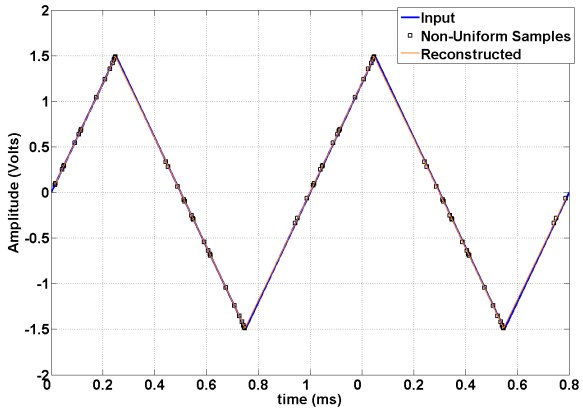


Figure 16: Reconstructed Linear

9.2 Electrocardiogram (ECG) Reconstruction

To test the new PDR data converter to acquire and reconstruct an analog signal from its nonuniform samples, a simulated electro-cardiogram (ECG) signal [25] was used. These signals have a wide dynamic range and "contain the QRS complex, which ensures oscillations near the Nyquist rate" [26] and are thus useful in exercising the nonuniform sampling architecture of the PDR.

The simulated ECG signal was modeled with a 1mV peak amplitude with 0.3mV DC offset and zero mean Gaussian noise with noise variance $\sigma_{\mathcal{N}o} \approx 0.058nW$ was added to the signal.

A wide view of the simulated run time of 10 seconds is shown in Figure 17, which demonstrates the PDR's ability to maintain the count. A zoom view of the reconstructed signal is shown in Figure 18, highlighting the features of the QRS pulse. When the signal is "idle" (DC like), and loiters near 0 Volts, the system continues to generate samples and is not adversely affected by the lack of signal dynamics.

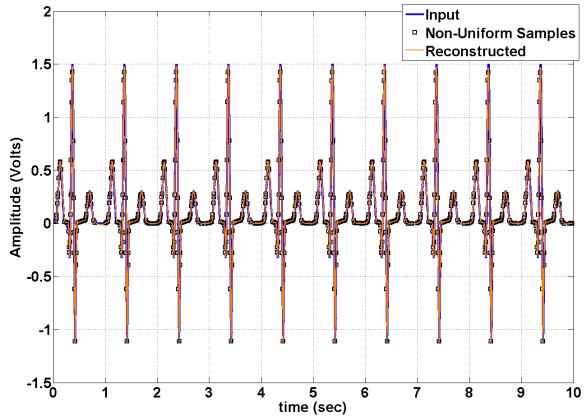


Figure 17: Reconstructed (simulated) ECG: Wide View

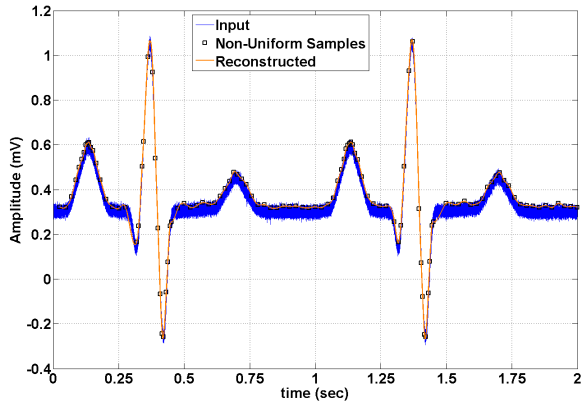


Figure 18: Reconstructed (simulated) Noisy ECG: Zoom View

10 Conclusion

A novel Analog-to-Digital Converter architecture based on partitioning the signal amplitude axis as a geometric series has been described. A detailed analysis of the design requirements to achieve the geometric partitioning has been provided and the essential circuits to realize the design presented. To extract the information content in each nonuniform digital sample, a proposed format of the nonuniform data was established, where it was shown that the partition number must be included in the digital word. Using reset Δ–PPM was shown to cause the system to satisfy the Nyquist requirement on average, and the geometric partitioning was shown to cause the SQNR to attempt to remain approximately constant. Lastly, the linearity of the PDR and the reconstruction of a simulated ECG signal were illustrated through simulation.

Conflict of Interest The authors declare no conflict of interest.

References

[1] C. Pappas, "A New Non-Uniform ADC: Parallel Digital Ramp Pulse Positino Modulation" in 31st IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Quebec Canada, May 2018. https://doi.org/10.1109/CCECE.2018.8447844

[2] S. Maymon, A. Oppenheim, "Sinc Interpolation of Nonuniform Samples" IEEE T SIGNAL PROCES, **59**(10), 4745–4758, 2011. https://doi.org/10.1109/TSP.2011.2160054

[3] D. Donoho, P. Stark, "Uncertainty Principles and Signal Recovery" SIAM J. Appl. Math., **49**(3), 906–931, 1989. https://doi.org/10.1137/0149053

[4] E. Candes, T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?" IEEE T INFORM THEORY, **52**(12), 5406–5425, 2006. https://doi.org/10.1109/TIT.2006.885507

[5] Y. Eldar, G. Kutyniok, Compressed Sensing Theory and Applications, Cambridge University Press,2012.

[6] F. Marvasti, Nonuniform Sampling Theory and Practice, Kluwer Academic/Plenum Publishers, 2001.

[7] J. Mark, T. Todd, "A Nonuniform Sampling Approach to Data Compression" IEEE T COMMUN, **29**(1), 24–32, 1981. https://doi.org/10.1109/TCOM.1981.1094872

[8] N. Sayiner, H. Sorensen, T. Viswanathan, "A Level-Crossing Sampling Scheme for A/D Conversion" IEEE T CIRCUITS SYST, **43**(4), 335–339, 1996. https://doi.org/10.1109/82.488288

[9] E. Allier, J, Goulier, G. Sicard, A. Dezzani, E. Andre, M. Renaudin, "A 120nm Low Power Asynchronous ADC" in ISLPED '05. Proceedings of the 2005 International Symposium on Low Power Electronics and Design, San Diego USA, 2005. https://doi.org/10.1145/1077603.1077619

[10] T. Wang, D. Wang, P. Hurst, B. Levey, S. Lewis, "A Level-Crossing Analog-to-Digital Converter with Triangular Dither" IEEE T CIRCUITS SYST, **56**(9), 2089–2099, 2009. https://doi.org/10.1109/TCSI.2008.2011586

[11] M. Malmirchegini, M. Kafashan, M. Ghassemian, F. Marvasti, "Non-uniform sampling based on an adaptive level-crossing scheme" IET SIGNAL PROCESS, **9**(6), 484–490, 2015. https://doi.org/10.1049/iet-spr.2014.0170

[12] S. Qaisar, M. Ben-Romdhane, O. Anwar, MTlili, A. Maalej, F. Rivet, C. Rebai, D. Dallet, "Time-domain characterization of a wireless ECG system event driven A/D converter" in 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Turin Italy, 2017. https://doi.org/10.1109/I2MTC.2017.7969682

[13] S. Naraghi, "Time-Based Analog to Digital Converters", Ph.D Thesis, University of Michigan, 2009.

[14] P. Maechler, N. Felber, A. Burg, "Random Sampling ADC for Sparse Spectrum Sensing" in 2011 19th European Signal Processing Conference, Barcelona Spain, 2011.

[15] S. Becker, "Practical Compressed Sensing: Modern Data Acquisition and Signal Processing", Ph.D Thesis, California Institute of Technology, 2011.

[16] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. Romberg, A. Emami-Neyestanak, E. Candes, "A Nonunifom Sampler for Wideband Spectrally-Sparse Environments" IEEE J EM SEL TOP C, **2**(3), 516–529, 2012. https://doi.org/10.1109/JETCAS.2012.2214635

[17] R. Gray, D. Neuhoff "Quantization" IEEE T INFORM THEORY, **44**(6), 2325–2383, 1998. https://doi.org/10.1109/18.720541

[18] W. Kestler (Editor), The Data Conversion Handbook, Analog Devices Inc., Elsevier-Newnes, 2005.

[19] C. E. Shannon, "Communication in the Presence of Noise" Proceedings of the IRE, **37**(1), 10–21, 1949. https://doi.org/10.1109/JRPROC.1949.232969

[20] H. Shapiro and Richard A. Silverman, "Alias-Free Sampling of Random Noise", New York University - Institute of Mathematical Sciences, 1959.

[21] F. Beutler, "Error-Free Recovery of Signals from Irregularly Spaced Samples" SIAM J. Appl. Math., **8**(3), 328–335, 1966. https://doi.org/10.1137/1008065

[22] R. Baker, H. Li, and D. Boyce, CMOS Circuit Design, Layout and Simulation, IEEE Press Series on Microelectronic Systems, Wiley Interscience, 1998.

[23] W.R. Bennett, "Spectra of Quantized Signals" Bell Sys. Tech. Journal, **27**(3), 446–472, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01340.x

[24] Subbarayan Pasupathy, *Minimum Shift Keying: A Spectrally Efficient Modulation*, IEEE Communications Magazine, Vol: 17, Issue: 4, July 1979

[25] Sophocles J. Orfanidis, Introduction to Signal Processing, Pearson Education/Prentice Hall, 1996.

[26] Marco A. Gurrola-Navarro, "Frequency-Domain Interpolation for Simultaneous Periodic Nonunifom Samples" in 2018 IEEE 9th Latin American Symposium on Circuits and Systems (LASCAS), Puerto Vallarta Mexico,, 2018. https://doi.org/10.1109/LASCAS.2018.8399937

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 1, 183-192 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Enhancing and Monitoring Patient Outcomes Through Customized Learning

Majed Almotairi, Mohammed Abdulkareem Alyami, Yeong-Tae Song*

Dept. of Computer & Information Sciences, Towson University, 21252, USA

ARTICLEINFO

Article history:
Received: 21 December, 2018
Accepted: 31 January, 2019
Online: 13 February, 2019

Keywords:
Patient Education
Customized learning
Dublin Core Metadata
Learning Object
Learning Object Repository

ABSTRACT

Chronic diseases such as heart disease, cancer, diabetes, and asthma continue to increase in the general public within the modern era. With careful observation of the symptoms potential diseases may be detected early and managed properly. For that to happen, the awareness of the symptoms and proper knowledge about the diseases may be needed for each patient. To acquire such knowledge, patients may need to gather essential health information from a variety of sources such as the Internet, articles, or some type of e-learning systems. However, the amount of available information, often too much, which discourages patients to continue. In response to such scenarios, we propose an approach that delivers only the relevant information that is specific to each patient's condition. In our approach, we utilized observed symptoms and vital signs to identify potential diseases of a patient. As they use the system, their profile may be constructed to deliver patient-specific set of learning materials called a study plan. To monitor and promote their study, we developed a mobile application that allows patients to view their study plan(s) and to study the customized learning materials. Such customized learning allows patients to take control of their symptoms and potential diseases, which eventually helps them to improve their outcomes as a result.

1 Introduction

Chronic diseases such as heart diseases, cancer, diabetes, and asthma are continuously increasing [1]. As reported by the Centers for Disease Control and Prevention (CDC) [2], 60% of adults in the US have one chronic disease, 40% have two or more chronic diseases, and almost 40% of adults were considered obese in 2015-2016. Annually, over 1.7 million people are diagnosed with cancer, and more than a third die from it. The US health care budget on cancer is incessantly increasing and anticipated to reach $174 billion by 2020. Chronic diseases require continuous treatment, which decreases the quality of life and increase medical expenses. According to the CDC [1], about 90% of The US health care expenditure is spent on individuals with chronic diseases and mental health conditions. Hence, chronic diseases are considered highly costly compared to other health problems. However, these diseases may be prevented or improved by early detection and proper management. According to many of the randomized controlled trials, educational techniques may help decrease pain, increase coping skills, and decrease the primary care visits, which saves time and money [3,4]. Patients also need to learn about how and when to take their prescription drugs, and other information such as side effects. If patients do not follow the instruction on prescription, this may lead to undesirable results such as the status of disease getting worse, or even death, which eventually increases health care budget in the United States [5]. Therefore, it may help improve patients' outcome when care givers ensure that patients have read and understood the drugs' instructions.

When patients suffer from some symptoms such as chest pain, they would look for the information related to that. However, getting the proper knowledge about that can be quite difficult as patients may be overwhelmed by the vast amount of information that is available from various sources. They do not know where to start to educate themselves, as shown

*Yeong-Tae Song, Email:ysong@towson.edu

https://dx.doi.org/10.25046/aj040118

in Fig. 1. CDC and WebMD, for instance, provide a vast amount of Internet accessible information regarding prescription drugs, diseases and their conditions, healthy living, public health statistics, and more that can be utilized by general public [2].

However, the efficient consumption of the available information by patients can be difficult due to the amounts and types of information. To overcome such issues, we propose an approach that filters out irrelevant information and provide only the necessary information for the patient, i.e., customized patient education. In our study, we deliver patient education by providing customized learning materials based on their prescription drugs, clinical conditions, and potential diseases. Therefore, only the relevant information that is specific to each patient is delivered. The learning materials are queried semantically and delivered to the patients from the cloud based learning object repository. To ensure effective learning, patients' levels of understanding in health information may be gained to control the difficulty level of learning materials. In particular, semantic queries can be performed based on a patient's clinical condition and learning needs. Then the obtained results from the queries are filtered and organized based on the patient's knowledge level and relevancy. Based on the query results, a study plan can be created and organized from the learning object repository, which satisfies the learners' needs. In addition, we employed an assessment technique to monitor the patient's level of understanding on the learning materials and to ensure that he/she has read and understood the content of the learning materials. The remainder of this paper is arranged as follows: Section 2 discusses the literature review. Section 3 gives an overview of the proposed Patient E- Learning System (PELS), discusses the role of the system components, and presents the implementation of the suggested system. Finally, we conclude our study in section 4.

2 Literature Review

In the modern era, there are numerous contributions and researches about e-learning systems in various domains. For example, In [6], the authors offered a virtual medical school as an e-learning system that provides a problem-based e-learning environment that enables medical students and residents to access clinical cases by using the Hospital Information System (HIS). Ouf et al. [7] have used ontology and Semantic Web Rule Language (SWRL) to develop a smart e- learning ecosystem that delivers the associated research articles from different sources, such as IEEE library, Science Direct and Springer Link, to learners according to their needs (for instance preferred learning activities). In addition, other studies concentrated on ubiquitous learning [8-10], language learning, sharing of learning resources and mobile learning [11,12]. Mesquita and Peres [13] offered a Customized x-Learning Environment model that delivers customized learning materials by allowing learning to take place anytime / anywhere, depending on the student's needs and characteristics. However, the majority of the studies mentioned above rely on the learner profile generally in order to improve learner's achievements. Additionally, there is a lack of conducted studies on utilizing e-learning systems for patients and clinicians to use. Consequently, the nucleus of our research is to develop learning management system for clinicians and patients to use in order to improve patients health knowledge with an aim to improve patients outcomes. Despite the fact that there are many available resources (e.g. websites and mobile applications) that are reachable online for example PatientsLikeMe, MAYO Clinic, WebMD, iTriage, and others, patients are overwhelmed by the amount of learning materials. For example, PatientsLikeMe enables patients with the similar disease(s) to share information regarding their treatments, medical experiences, medication, outcomes, and more. However, room for medical errors exists while using this site since the patient is the one who generates the data. There are several reasons why these medical errors may take place, amongst them misconstruction due to limitation of using medical standards (e.g. patients may use of diverse terms to describe their problem), lack of using other factors that might be related to disease such as vital signs and demographic data, and concentration on diseases only [14]. In contrast, in our study, we utilized medical standards (e.g. RxNorm, SNOMED CT and ICD), vital signs and demographic data in order to provide learning materials to the patient according to his/her prescription drugs, diseases and symptoms and eventually avoid medical errors. In our study, an innovative learning approach is proposed by using customized learning. Based on this, a learning model is designed to achieve personalized learning experiences for patients. We can achieve this by retrieving the most proper learning objects and then providing them to proper patients. This, in turn, may encourage patients to educate themselves and reduce the time of seeking such learning materials from different resources.

2.1 Related Standards

This section describes the most common related standards that are available in the industry, and how they can be utilized in the PELS.

2.1.1 Patient Profile

In the modern era, there are various learner profile standards, including, but not limited to, Felder-Silverman Model, Kolb's Experiential Learning Model and Myers-Briggs Type Indicator, each individual standard is suitable for a specific learning domain [15]. However, Luciana et al. [16], and Ioannis et al. [17] reported that the two widely held standards that are utilized to define the learner profile are: Public and Private Information (PAPI), introduced by Institute of Electrical and Electronics Engineers (IEEE) [18], and Learner Information Package (LIP), introduced

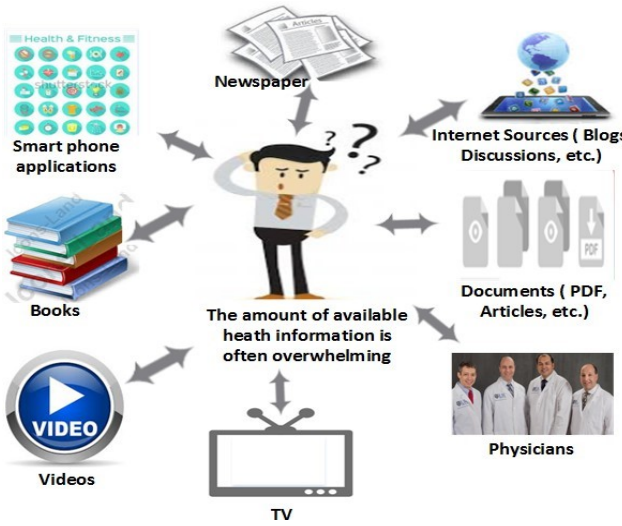


Figure 1: Patients are overwhelmed by the amount of available information

by Instructional Management System (IMS) [19]. The PAPI model contains six core classifications (personal information,security information, performance information, relation information , preferences information,and portfolio information) [18]. The IMS LIP model contains 11 core classifications (identification, affiliation, interest, accessibility, relationship, competency, activity, QCL [qualifications, certificates and license], goals, transcript, and security key) [19]. In addition, other extensions are available in each classification of these two standards that enable dealing with more possible characteristics [18,19]. Evangelou et al. [20] expresses that the PAPI model is generic and was developed to be utilized by variety types of systems or applications, but it does not afford information about dynamic characteristics of the learner profile. However, LIP is useful in the case that the system needs to be interoperable with other systems [20]. As mentioned above, the PAPI and LIP models comprise different categories to describe the learner profile, but none of these models entirely meet the PELS requirements, such as allowing clinicians to input their patients' symptoms and vital signs regularly. For that to happen, we developed a patient profile that involves specific characteristics to describe both dynamic and static information for each patient, which are utilized to gather patients' information in order to deliver customized learning materials to each individual.

2.1.2 Learning Object

A variety of learning object (LO) metadata standards are available in industry that are utilized to describe the LOs, which in turn, enables the ease of discovery and retrieval of LOs. An example of these standards, Learning Object Discovery and Exchange (LODE), was introduced by IMS Global (www.imsglobal.org). Another example of the standards, is called the Learning Object Metadata (LOM) developed by IEEE, which offers a technique to describe the profile of the LOs for use in e-learning systems [21]. Furthermore, one of the most common standards that is available is called the Dublin Core (DC), which also can be utilized to represent the LO profiles [22]. For generating and packaging the LOs, a well-known standard is called Common Cartridge (CC), which offers a standardized technique to create the LO content [23]. The CC also provides a standardized way to make interoperable e-learning systems and learning materials [23]. The CC utilizes the core components of the DC to describe the LO [22]. In our research, we utilized the CC standards that best fit our purpose. Most of the well-known e-learning systems use the CC to create the LOs as it offers standard methods to create LOs.

2.1.3 Medical Standards

Alyami et al. [24] summarizes the medical standards that are used at the present time. The ones that are related to our study are:

- Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT): a comprehensive standardized clinical terminology [24].
- International Classification of Diseases (ICD-9/ICD-10): used for epidemiology, health management and clinical purposes [24].
- RxNorm: used to provide normalized names for clinical drugs [24].

These three standards are used in the PELS for the clinicians use when they input data in the patients' profile (e.g. symptoms and prescription drugs).

2.1.4 Function Modeling Notations:

The US Air Force Program for Integrated Computer Aided Manufacturing (ICAM) presented various modeling notations that they have developed to provide better analyses and communication techniques in order to increase the manufacturing efficiency [25]. The main modeling notation that ICAM presented is called IDEF0 (ICAM Definition), which is based on Structured Analysis and Design Technique TM (SADTTM) [25]. IDEF0 is used to create a functional model that represents the functions, activities and processes that are involved in a system [25].

2.2 Limitations of Existing E-learning Systems

We examined the contemporary e-learning systems that are available for the public in industry, and the results show that there are more than 400 e-learning systems, either commercial, free or open source systems. In order to obtain functions that can be utilized for clinical education appropriately, we utilized the main functions from the free and open source well-known learning management systems, including Moodle, ATutor, Dokeos, ILIAS and Sakai. In addition, we took into

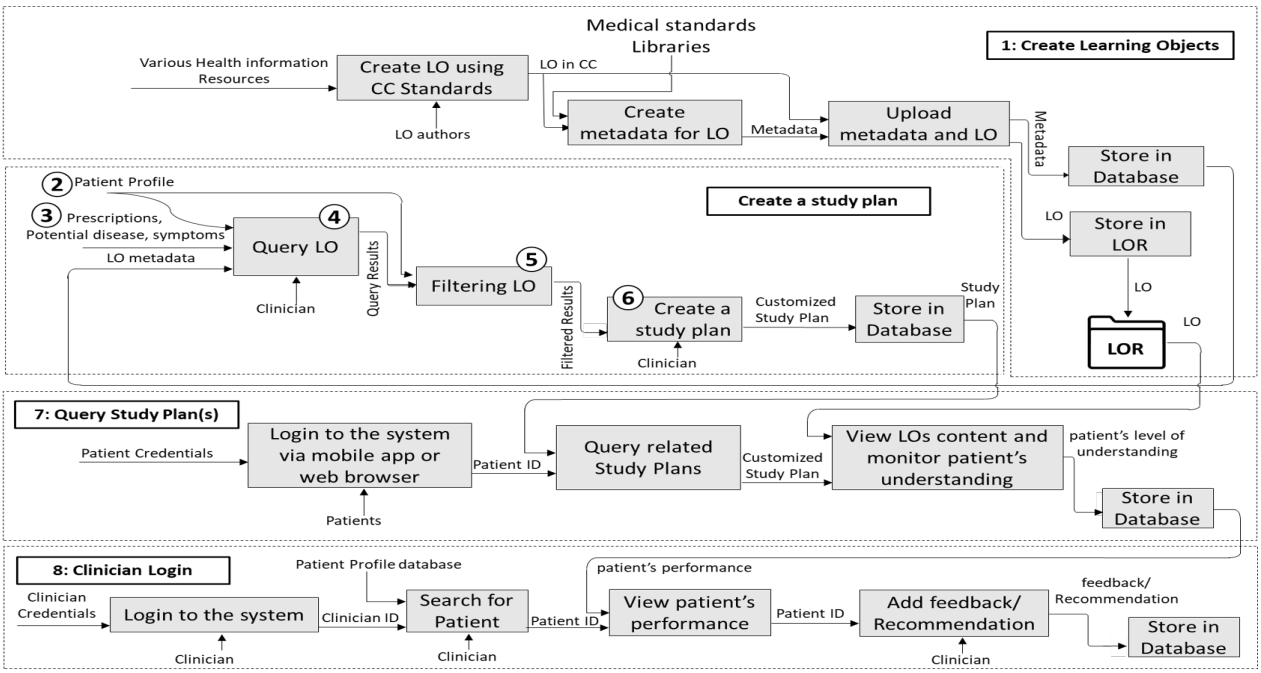


Figure 2: Main Functions of the Patient E-learning System (Numbers: 1, 2, etc. in this figure correspond to section 3)

account the other organizations that offer information for patients, such as CDC, WebMD and MAYO Clinic. The results indicate that there are many limitations of these systems, including:

- In the most of the e-learning systems, the instructor chooses particular content to be accessible by learners, which make the system a teacher-centered instruction.
  - Limited options are available for customized learning or filtering for learning materials that can be utilized to satisfy learners' needs and goals. However, patients are required to use the search function within these e-learning systems to query the desired information.
  - Enrollment is required in most cases by each individual learner in a specific course to be able to contribute and view the course materials where the instructor is the person who evaluates the students' performance.
- Limited development of learner profiles that may allow for saving and analyzing the current learner's condition in order to offer learning materials based on their needs.
- MAYO and WebMD organizations have developed web applications that deliver clinical information to the users according to their inputs (e.g. prescription drugs and symptoms). However, they do not provide any assessment techniques to ensure that users have read and understood the content. In addition, they do not enable patients to save their information so they can monitor the changes of the patient outcomes.

To cope with these boundaries and limitations, we proposed new functions and improved some of the functions of existing applications to be used in the PELS as follows:

- To generate the learning object, we followed the CC standard for the creation of the learning objects, and we used the DC metadata to represent the learning object in order to facilitate the finding and retrieval of learning materials that are associated with the patients' education.
- As none of the current learner profiles that exist in the current e-learning systems are appropriate for the patient education, we offered a patient profile with specified characteristics that can deal with static and dynamic information, such as patient's vital signs and symptoms. We utilized patient profiles to enable the following:
  - To allow clinicians to save and update patients' information (e.g. symptoms and potential disease[s]).
  - Dynamic retrieval of customized learning materials that meet patients' needs according to the available information in patients' profiles. Therefore, our proposed system is not similar to those that ask users to type keywords in the search textbox to retrieve their desired learning materials.
- Assessment techniques are conducted to ensure that patients have read and understand the learning materials.

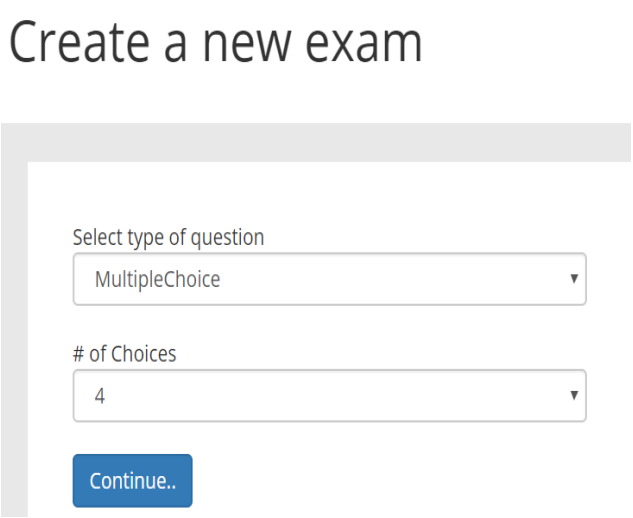


Figure 3: Exam creator tool for LO authors

- For the ease of use, we developed a mobile application that allows patients to view their learning materials straightforwardly.
- Although the PELS is an independent system, well-known standards (e.g. medical and metadata standards) were utilized to record such data in our proposed system. Therefore, the data can be integrated with other systems if they utilize the same standards.

3 The Proposed Patient E-Learning System (PELS)

Patients and clinicians are the main users in our proposed PELS. For that reason, the limitations of the systems that provide information for patients and the current e-learning systems are taken into account. In addition to the functionality of our proposed system that was presented in [1], we also proposed more enhancements by adding more functions that are used to deliver learning materials to patients based on their need and capture their performance in order to insure that patients have learned the desired knowledge. To represent the functionality of the PELS, the IDEF0 notation [25] has been used to show the main components of the proposed system along with the main input(s) and output(s) of each component as shown in Figure 2. The following sections illustrate in detail how each component is used in the PELS.

3.1 Creating Learning Objects

We utilized the CC standard [23] in the process of creating the LOs which make the PELS organize and retrieve the LOs efficiently. In addition, the CC standard utilizes the DC standard for the metadata representation. The processes of creating LO are shown in the Figure 2 section 1.First, the LO authors create LO using

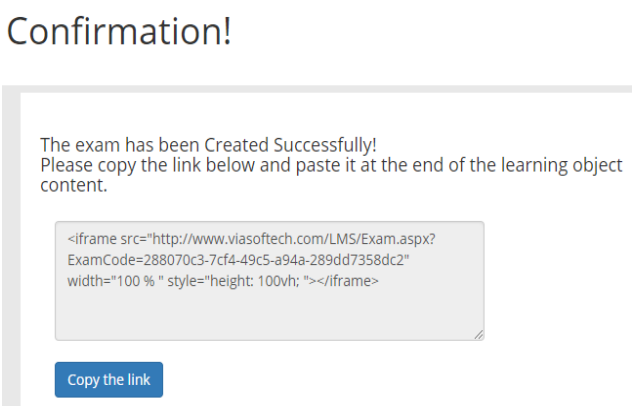


Figure 4: Example of the generated embed code

any content authoring tool that create LO in CC standards. While creating the LO, most of the content authoring tools allow authors to create exams and quizzes for each LO to ensure that learners have read and understood the content of the LO.

However, by using these authoring tools, it is difficult for the PELS to capture and save patient's answers, as there is no option to integrate LOs with the PELS (or third party systems in general) that allow developers to capture learners' inputs and store them in a database. Therefore, we developed an exam creator tool that allows LO authors to create an exam that can be managed by each clinician to capture their patients' performance. However, patients may not be able to answer questions that require typing, therefore, we only focused on multiple-choice and True/False questions that are easy for patients to answer and for the clinician to receive the final score without reading through the responses.

To create an exam, the LO authors create a title for the exam as first step, and then they choose the type of question(s) which can be either multiple-choice or True/False. If the author chooses a multiple-choice question, another menu appears that allows them to specify the number of choices as shown in Figure 3. The LO author can repeat this step several times to add the desired number of questions they want. When they finish adding questions, the exam creator tool generates an embed code that allow LO authors to embed the exam with the LO as shown in Figure 4. Once LO authors complete the process of creation the LO, they need to add metadata based on the DC standard to represent the LO. DC consists of fifteen core elements that can be used to define the LO profile more precisely [22]. In this research, we used the DC metadata elements to describe clinical LOs efficiently, as shown in Table 1. The LO authors are required to fill most of these elements in order to gain metadata records that can be queried efficiently. An example of the learning objects' metadata that are stored in the database is shown in Figure 5. The World Health Organization (WHO), that provides International Classification of Diseases (ICD-9/ICD- 10), grouped the potential diseases based on the body systems and other factors. In our study, we

Table 1: Metadata schema for the LO

Element	Description
Title	Learning object's title
Creator	Learning object's author
Subject	Keywords that outlines the LO
Description	Detailed description about the LO
Relation	The related medical standard code(s) that is related to the topic of the LO, which also can be used to indicate whether the LO is specific or generic
Date	The creation date of the learning object
Type	The category of the LO (e.g. symptoms, causes, diagnosis and drugs)
Identifier	The ID of the learning material
Language	The spoken/written language used in the content of the LO (e.g. English and Spanish)
Rights	Access rights and copyright of the LO
Source	The source of the LO

utilized these classifications to determine whether the LO is generic (e.g. body system) or specific (e.g. specific disease).

The PELS utilizes the Learning Object Repository (LOR), which is a directory that is used to organize and store all learning materials. Accordingly, once the LO authors complete filling the metadata of a LO and click "upload metadata and LO", the system creates a separate sub-directory inside the LOR for each new LO with a unique label and adds this unique label to the LO's metadata to be efficiently and effectively queried.

The content of the LOs can be collected from various recognized health information resources. However, the creation of the LO can be collected manually using any e-learning content authoring tool that generates LO(s) in CC standard. In addition, other organizational websites (e.g. CDC [2]) provide Syndication API, that collect an enormous number of learning materials from their repository. All learning materials are stored and organized in the LOR to allow the PELS to use them and retrieve appropriate ones for the patient efficiently. During the exploratory analysis, we collected content for variety types of diseases, symptoms and drugs from different well-known health information resources such as CDC [2] and created LOs based on the CC standard.

3.2 Patient Profile

The Patient profile is developed to store patients' information regularly. As patients' information may include clinical information such as prescriptions, symptoms and vital signs, clinicians are involved in the process of adding the clinical information to the patient profile. In this study, we categorized the patient profile into two groups: dynamic (clinical information) and static (demographic information) information for each patient. The PELS utilizes patient profiles to gather information about each patient in order to deliver customized learning materials to each individual patient. The components of the patient profile can be listed as

Id	Title	Description	Lang...
1	Genital Warts | S...	Genital warts ...	EN
2	How to prevent ...	Pertussis (wh...	EN
3	Vitiligo: Get the F...	Vitiligo is a co...	EN
4	Strep Throat: Sy...	Strep throat is...	EN
5	Key Facts About ...	Influenza (als...	EN
6	Pink Eye Facts: I...	"Pink eye" is ...	EN
7	The Best and W...	Picking the ri...	EN
8	Acute Maxillary S...	Acute sinusiti...	EN
9	Scabies: Images,...	Scabies is a v...	EN
10	What Causes Kid...	Kidney stones...	EN

Figure 5: Portion of the learning objects metadata

follows (see Figure 6):

- Patient demographic information: consists of full name, date of birth, gender, address, username, password, email address and phone number.
- Patient level of knowledge and preference: Level of knowledge is used to determine the learning materials' difficulty level that patient can understand. Preference contains content types such as text, video, etc., preferred language and input / output device.
- Patient vital signs: updated frequently, which includes respiratory rate, pulse, temperature, blood oxygen saturation and blood pressure.
- Patient symptoms: includes symptoms that patients observe, which may be used to determine the potential disease(s). Potential diseases: includes disease(s) that patients may have based on their symptoms.
- Prescriptions: includes prescription drugs that patients obtained.
- Study Plan: a collection of customized LOs for each individual patient to study.

On the other hand, clinicians' profile consists of the following:

- Demographic information: consists of full name, date of birth, gender, address, username, password, email address and phone number.
- Patient list: a list of patients that the clinician deals with. In order for the clinician to add a patient in their list, the patient's approval is required.

By using PELS, a patient signs up as a new patient and starts adding his/her main information and preferences (static information). However, patients may not be able to deal with clinical information. Therefore,

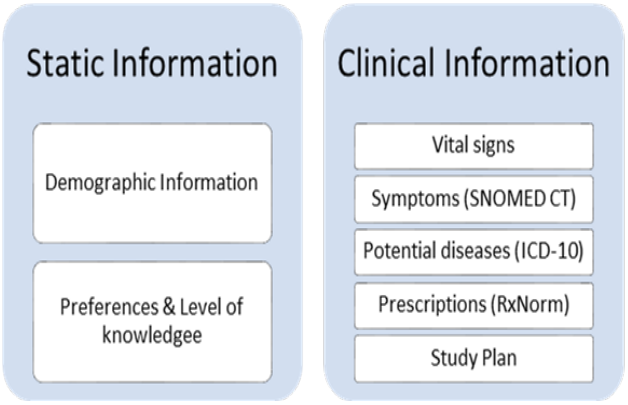
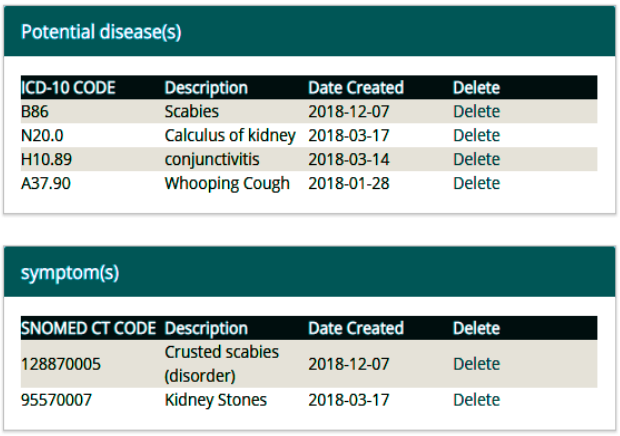


Figure 6: Components of the Patient Profile



Potential disease(s)

ICD-10 CODE	Description	Date Created	Delete
B86	Scabies	2018-12-07	Delete
N20.0	Calculus of kidney	2018-03-17	Delete
H10.89	conjunctivitis	2018-03-14	Delete
A37.90	Whooping Cough	2018-01-28	Delete

symptom(s)

SNOMED CT CODE	Description	Date Created	Delete
128870005	Crusted scabies (disorder)	2018-12-07	Delete
95570007	Kidney Stones	2018-03-17	Delete

Figure 7: List of potential diseases and symptoms in the patient profile

the PELS limits adding clinical information only to clinicians. Consequently, a clinician needs to search for a patient first using search criteria, and then sends a request to the patient to add him/her to their list. The patient then will have the option to approve/reject the clinician's request. Once the patient approves the clinician's request, the clinician then can add clinical information to the patient profile.

3.3 Potential Diseases and Prescriptions

With the aim of creating study plans, the PELS needs to determine the potential diseases that patients may have, observed symptoms that they are suffering from, or prescription drugs that clinician prescribed. To collect such data, potential diseases can be entered in two methods: directly and indirectly. From the clinician directly, if the clinician already know the (potential) disease that patient suffers from. Indirectly, which can be obtained by using the Health Decision Support System (HDSS) that we have developed [24] which relies on medical standards such as SNOMED CT and ICD-9/ICD-10. HDSS assists patients and clinicians in identifying such potential disease(s) based on symptom(s) they provided. For instance, if a patient utilizes HDSS to determine his/her potential disease and enters his/her symptoms through the system, PELS can store this information on the patient profile, as shown in Figure 7. On the other hand, the prescription drug(s) need to be determined in the patient profile in order to educate patients about how to use specific drug, side effects when using this drug, and more. The PELS utilizes the RxNorm standard that allows clinicians to utilize standard terms for clinical drugs when adding them to the patient profile.

3.4 Query Module

As patients may not know how to deal with clinical terms or what is the right learning materials they should start with, clinicians are involved in this process to make sure that patients receive learning materials based on their needs. To query the metadata, Microsoft offers the semantic search that built upon the full-text search feature in the SQL server [26]. It extracts and statistically indexes the relevant key phrases. These key phrases are used to find and index the similar or related results [26].

The SQL server provides mechanisms that are utilized to implement the full-text search, for instance [26]:

- Stop list: deals with list of words commonly taking place in strings that do not help the search which should to ignored when performing a query e.g. "in" or "an".
- Stemmer: deals with the inflectional forms of a word, for instance, in the word "play" the full-text searches for the stems of the word, such as played, playing and player.
- Thesaurus: deals with the synonyms of a word, for example bike and bicycle.
- Word breaker: deals with the word boundaries as in the words "multi-millions" and "multimillions".
- Replacements: deals with the most frequently misspelled words such as Calendar/ Calender.

The clinician manages this process in order to ensure that patient receives the desired learning materials. In this way, amongst variety kinds of LOs that are stored in the LOR, and based on the available information in the patient profile, the PELS analyzes this information through the query module in order to deliver the proper learning materials to the patient. To do so, this module systematically queries the most associated learning materials that the patient needs based on his/her conditions. The query results (retrieved LOs) go through to the filtering module.

3.5 Filtering Module

The filtering module gets called after the query module is executed in order to retrieve customized learning materials based on the patient's profile. It filters the learning materials according to the patient's

Learning Object Title	View Contents
How to prevent Pertussis (Whooping Cough) (Not Completed!) This Learning material is recommended to you because **it is associated with the potential disease Whooping Cough** that is listed in your profile	View Contents
Pink Eye Facts: Identify Symptoms and Treat Pink Eye (Conjunctivitis) (New!) This Learning material is recommended to you because **it is associated with the potential disease Conjunctivitis** that is listed in your profile	View Contents
Key Facts About Influenza (Flu) | Seasonal Influenza (Flu) (Completed!) This Learning material is recommended to you because **it is associated with the potential disease Influenza (disorder)** that is listed in your profile	View Contents

Figure 8: Customized study plan based on patient profile

preferences and/or the level of knowledge. It also eliminates learning materials that the patient has already completed in the past. In this manner, the module retrieves only the related learning materials for an individual patient, i.e., customized learning contents, to the patient.Figure 5 and Figure 8 show the differences before and after filtering.

3.6 Creating Study Plans

The study plan is a collection of learning materials that patients can study to educate themselves about a certain topic [27]. In the PELS, the clinician is involved in this process to verify the process of selecting learning materials. The created study plan module automatically generates a collection of learning materials according to the results received from the query and filtering modules in order to provide it to the patient. Then clinicians can verify the choices of learning materials as they have the authorization of adding/deleting learning materials from the study plan. When the clinician confirms the creation of the study plan, the PELS generates a list of recommended learning materials as an XML that contains information about the recommended learning materials, and it points out the reason of recommending each learning material.Figure 8 shows how the patients view the content of the study plan in PELS. The PELS allows multiple creations of study plans. Therefore, each time that clinicians update patients' conditions (e.g. add new potential disease), this module allows clinicians to generate a new study plan that contains a new list of learning materials in order to provide efficient learning materials to each individual patient according to their up-to-date conditions.

3.7 Query Study Plan

This module is about how the PELS queries the study plan(s) and delivers them to the right patient. Patients may not like the way that they need to use the personal computer/laptop to be able to sign in to the PELS and study. Therefore, we make the use of the

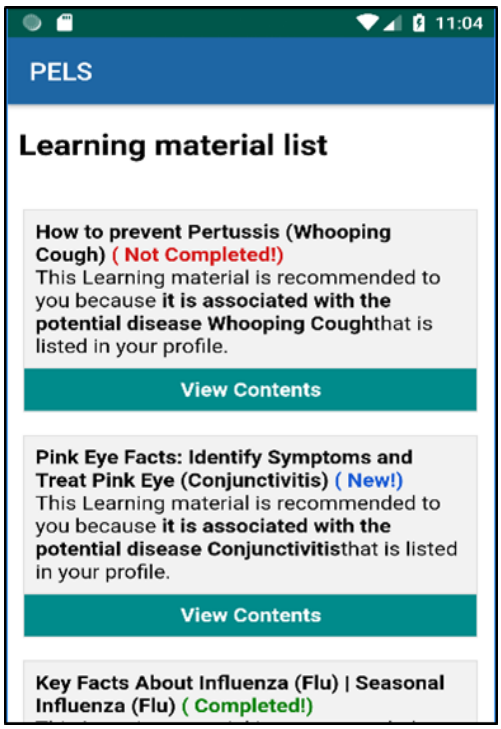


Figure 9: View study plan content through the mobile application

PELS even easier by developing a mobile application that allows them to view and navigate through their study plan(s) and display the content of each study plan. Figure 9 shows how patients view study plan(s) when they use the mobile application. In order for the PELS to query the patient's study plan(s), the patient first needs to log in using their credentials. Once the patient logs in successfully, the PELS queries the list of study plan(s) associated with this particular patient. Each study plan includes one or more learning material(s) that are approved by a clinician as shown in Figure 8. In order to monitor how patients consume the learning materials, the PELS categorizes three different statuses that appear next to the title of the learning object as shown in Figure 8. The time when each of these three statuses that appear can be illustrated as follows:

- New: appears when the patient still has not viewed the learning material yet.
- Not Completed: each learning material contains a quiz that needs to be completed by the patient to ensure that he/she has read and understand the content of the learning material. This status appears when the patient viewed the learning material but still has not attempt to complete the quiz.
- Completed: appears when the patient spent some time to reading the learning material and has completed answering the quiz.

If the status of the learning material remains new or not completed for a certain period of time (e.g. a week),

the system sends a reminder notification to the patient saying that the learning material needs to be completed. However, at the end of each learning material

Figure 10: Patient's performance from the clinician's view

content, there is a quiz, which has been included to encourage patients to read the learning materials and make sure that they understand its content. Therefore, the PELS allows patients to take the quiz multiple times if needed. Accordingly, when the patient chooses the wrong answer, the PELS shows a hint to tell the patient that he/she needs to change their response and find the correct answer. On the other hand, when the patient views the content of the learning material, the PELS contains quantitative measures that capture the patient's performance of metrics, for instance:

- Total time that the patient spent on reading the learning material.
- Total number of attempts that the patient made to complete the quiz.
- Total points scored on each attempt.

All the captured data are stored in the patient profile so the patient can see how he/she performed on each task. In addition, it can reachable by the clinician so they can see how their patient performed in completing their tasks (see next section).

3.8 Clinician Login

Clinicians need to have their own access to the PELS so they can perform particular tasks, such as creating study plans and monitoring their patients' performance. Therefore, the PELS provides certain views with their functionality that perform when the user signs in as a clinician. Accordingly, clinicians start signing in to the PELS using their credentials if they have already created their own account. If they want to deal with a new patient, they need to search for the patient using the search criteria and send a request to the patient to add him/her to their list. If the patient approves the request, the patient will be added to the clinician list. Accordingly, the clinician can select any patient from their list in order to perform any task

Time spent on reading the Learning Material	00:05:06
Total attempts to complete the quiz	2 attempt(s)

Attempt # 1

Submission Time	11/17/2018 9:19:27 PM
Total points scored	2 / out of 8

Attempt # 2

Submission Time	11/17/2018 9:34:11 PM
Total points scored	8 / out of 8

Figure 11: Patient's quiz result

such as adding a new potential disease(s) and recommending study plan(s). In addition, clinicians can view patients' performance and see how they performed in a specific recommended study plan as shown in Figure 10. The PELS shows the status of each learning material whether or not the patient has read the learning material, has read the learning material but still has not completed the quiz or if the patient has completed reading the learning material and has answered the quiz. When the status is completed, the PELS shows an option to view the results, which includes the metrics that are mentioned in the previous section, shown in Figure 11. However, the PELS has an optional feature that allows clinicians to add feedback/recommendation for their patient when needed.

4 Conclusion

In this paper, we provide customized e-learning for patients with the aim of increasing their level of health knowledge and eventually improving their outcomes. For this to happen, we analyzed patient health data, such as observed symptoms and vital signs, to identify potential disease(s) for the given health data. Such factors are utilized to build patient profiles (such as patient preferences) that are used to provide only the relevant learning materials. Our proposed system, PELS, showed how to gather and organize heterogeneous learning materials using CC and DC metadata. To facilitate the retrieval of relevant LOs, the DC tags are utilized to retrieve such learning materials. In addition, we use an assessment technique to ensure that patients have read and understood the content of the learning materials. Moreover, we enhanced the usability of the PELS by developing a mobile application to facilitate the ease of use. Through this approach, we provide customized education that may help to improve patient outcomes . For future work, we plan to utilize an ontology-based patient profile analysis, which may increase the level of accuracy and relevancy of the search.

References

[1] This paper is an extension of work originally presented in 2018 IEEE/ACIS 19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD) Busan, 2018. http://dx.doi.org/10.1109/SNPD.2018.8441054.

[2] CDC, "Centers for disease control and prevention," https://www.cdc.gov/, 2006, accessed: 2017-01-21

[3] K. Jordan, N. Arden, M. Doherty, B. Bannwarth, J. Bijlsma, P. Dieppe, K. Gunther, H. Hauselmann, G. Herrero-Beaumont, P. Kaklamanis et al., "Eular recommendations 2003: an evidence based approach to the management of knee osteoarthritis: Report of a task force of the standing committee for international clinical studies including therapeutic trials (escisit)," Annals of the rheumatic diseases, vol. 62, no. 12, pp. 1145 1155, 2003. doi:10.1136/ard.2003.011742.

[4] E. Superio-Cabuslay, M. M. Ward, K. R. Lorig et al., "Patient education interventions in osteoarthritis and rheumatoid arthritis: a meta-analytic comparison with nonsteroidal antiinflammatory drug treatment," Arthri- tis Care and Research, vol. 9, no. 4, pp. 292301, 1996. DOI: 10.1002/1529-0131(199608)9:43.0.CO;2-4.

[5] M. L. McCarthy, R. Ding, N. K. Roderer, D. M. Steinwachs, M. J. Ortmann, J. C. Pham, E. S. Bessman, G. D. Kelen, W. Atha and R. Retezar, "Does providing prescription information or services improve medication adherence among patients discharged from the emergency department? A randomized controlled trial," Annals of emergency medicine, vol. 62, no. 3, pp. 212–223, 2013. DOI: 10.1016/j.annemergmed.2013.02.002.

[6] F.-M. Shyu, Y.-F. Liang, W.-T. A. Hsu, J.-J. Luh, and H.-S. Chen, "A problem-based e-learning prototype system for clinical medical education," in Medinfo, 2004, pp. 983987. DOI: 10.3233/978-1-60750-949-3-983.

[7] S. Ouf, M. A. Ellatif, S. E. Salama, and Y. Helmy, "A proposed paradigm for smart learning environment based on semantic web," Computers in Human Behavior, vol. 72, pp. 796818, 2017. https://doi.org/10.1016/j.chb.2016.08.030

[8] S. Yu, X. Yang, G. Cheng, and M. Wang, "From learning object to learning cell: A resource organization model for ubiquitous learning," Journal of Educational Technology & Society, vol. 18, no. 2, p. 206, 2015.

[9] G.-J. Hwang, C. Hui-Chun, S. Ju-Ling, S.-H. Huang, and T. Chin-Chung, "A decision-tree-oriented guidance mechanism for conducting nature science observation activities in a context-aware ubiquitous learning environment," Journal of Educational Technology & Society, vol. 13, no. 2, p. 53, 2010.

[10] Y.-M. Huang and P.-S. Chiu, The effectiveness of the meaningful learning-based evaluation for different achieving students in a ubiquitous learning context, Computers & Education, vol. 87, pp. 243253, 2015. https://doi.org/10.1016/j.compedu.2015.06.009.

[11] Y.-J. Lan and Y.-T. Lin, "Mobile seamless technology enhanced csl oral communication." Journal of Educational Technology & Society, vol. 19, no. 3, 2016.

[12] J. Abramson, M. Dawson, and J. Stevens, "An examination of the prior use of e-learning within an extended technology acceptance model and the factors that influence the behavioral intention of users to use m- learning," SAGE Open, vol. 5, no. 4, p. 2158244015621114, 2015. https://doi.org/10.1177%2F2158244015621114

[13] F. Moreira, A. Mesquita, and P. Peres, "Customized x-learning environ- ment: Social networks & knowledge-sharing tools," Procedia Computer Science, vol. 121, pp. 178185, 2017. https://doi.org/10.1016/j.procs.2017.11.025

[14] J. H. Frost and M. P. Massagli, "Social uses of personal health infor- mation within patientslikeme, an online patient community: what can happen when patients have access to one anothers data," Journal of medical Internet research, vol. 10, no. 3, 2008. https://dx.doi.org/10.2196%2Fjmir.1053

[15] R. M. Felder and R. Brent, "Understanding student differences," Journal of engineering education, vol. 94, no. 1, pp. 5772, 2005. https://doi.org/10.1002/j.2168-9830.2005.tb00829.x

[16] L. A. Zaina, J. F. Rodrigues Jr, and G. Bressan, "An approach to design the student interaction based on the recommendation of e-learning objects," in Proceedings of the 28th ACM International Conference on Design of Communication. ACM, 2010, pp. 223228. https://doi.org/10.1145/1878450.1878488

[17] I. Panagiotopoulos, A. Kalou, C. Pierrakeas, and A. Kameas, "An ontology-based model for student representation in intelligent tutoring systems for distance learning," in IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, 2012, pp. 296305. https://doi.org/10.1007/978-3-642-33409-2_31

[18] IEEE, "IEEE standard for learning technology - public and private information (papi) for learners," http://metadata-standards.org/, 2002, accessed: 2017-09-04.

[19] IMS Global Learning Consortium Inc, "Ims learner information packaging infomation model specification," http://www.imsglobal.org/profiles/lipinfo01.html, 2001, accessed: 2017-09-07.

[20] C. E. Evangelou, M. Tzagarakis, N. Karousos, G. Gkotsis, and D. Nou- sia, "Augmenting collaboration with personalization services," Inter- national Journal of Web-Based Learning and Teaching Technologies (IJWLTT), vol. 2, no. 3, pp. 7789, 2007. DOI: 10.4018/jwltt.2007070105

[21] IEEE, "IEEE standard for learning object metadata," https://standards.ieee.org/findstds/standard/1484.12.1-2002.html, accessed: 2017-10-09.

[22] DCMI, "innovation in metadata design, implementation & best prac- tices," http://www.dublincore.org/documents/dces, accessed: 2017-10- 02.

[23] IMS Global, "Ims common cartridge specification," http://www.imsglobal.org/activity/common-cartridge, 2008, accessed: 2017-11-12.

[24] M. A. Alyami, M. Almotairi, A. R. Yataco, and Y.-T. Song, "Health decision support system based on patient provided data for both patients and physicians use," Presented in ACM IMCOM 2018,12th International Conference on Ubiquitous Information Management Information Man- agement and Communication. https://doi.org/10.1145/3164541.3164632

[25] F. PUBS, "Announcing the standard for integration definition for func- tion modeling (idef0)," Draft Federal Information Processing Standards Publication, vol. 183, 1993.

[26] Microsoft, "Semantic search (sql server)," https://docs.microsoft.com/en- us/sql/relational-databases/search/semantic-search-sql-server, accessed: 2017-06-12.

[27] Y.-T. Song, K.-e. Park, and Y. Yoon, "Ontology based learner-centered smart e-learning system," in Proceedings of VI International GUIDE Conference, Athens, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 190-196 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Nonlinear Analytic Modeling for Novel Linear Variable Reluctance Motors

Aymen Lachheb[*,1], Jalel Khediri[2], Lilia El Amraoui[1]

[1]*SEICT, LR18ES44, Enicarthage, 2035, Tunisia*

[2]*SIME, University of Tunis, 1008, Tunisia*

ARTICLE INFO

Article history:
Received: 05 January, 2019
Accepted: 03 March, 2019
Online: 26 March, 2019

Keywords:
Linear motor
Analytic modeling
Finite elements method
Fourier series

ABSTRACT

In this paper, an analytical modeling is proposed to compute the static and dynamic characteristics of linear variable reluctance motors with taking account of the saturation and the non-linearity of the magnetic circuit. The proposed model is based on the Fourier series expression of the phase flux in which the variation of the linkage flux with phase current is defined by arc tangent function. Then, an analytical expression for the co-energy and electromagnetic forces are derivative. The effectiveness of the proposed method is proved in terms of accuracy by comparing the computed results obtained by the two dimensional finite elements method with analytic model. Subsequently, the proposed modeling approach is simple and provides accurate representation of linear variable reluctance motor, as confirmed by the compared results.

1. Introduction

Linear variable reluctance motor (LVRM) becomes one of the most promising candidates for direct drive systems [1] and it will be more attractive because of its higher developed force and its immense efficiency which make it more competitive with conventional rotary motor [2]. Furthermore, the LVRM is highly robust with low cost manufacturing, this allows linear motors to be used in the numerical control of machine tools and in robotics applications. Hence, a drive system with LVRM has gained a great interest for several industrial application [3].

However, the LVRMs operate with a high magnetic flux density in order to produce a maximum of power, which causes a saturation in the magnetic circuit. In addition, the characteristic of LVRM is strongly dependent on the state of the magnetic circuit, which is depending on the supplying current of windings and the position of translator relative to the stator teeth [4]. Then, the performance of LVRM is strongly influenced by the saturation of magnetic circuit for high levels of flux density, this phenomenon is repeated periodically as long as the translator move out from the unaligned position to reach the aligned position. In fact, the magnetic non–linearity makes the study of this motor more difficult.

Therefore, to predict the static and the dynamic performance of the motor precisely requires a correct representation of magnetic quantities, it is necessary to elaborate an analytical model, which takes account of the saturation effects in the magnetic circuits during the operation of this motor.

Several modeling technique has been proposed recently for studying the static and dynamic performances of the linear motor, finite element method (FEM) is one of popular approach adopted for the non-linear magnetic analysis of electrical machines [5], but it requires high computer performance and the calculation time needed to solve the equations governing the motor operation is quite long. For this reason, that we find many research works are adopting the analytic modeling to determine the characteristics of the variable reluctance motor [6].

Recently several research works have dealt with modeling and simulation of switched reluctance motor, some models are based on the expression of the inductance on the other hand, some model is based on the expression of the flux linkage [7, 8].

In reference [9], a nonlinear model is proposed which takes into account the magnetic saturation, a nonlinear representation of the phase inductance has been used to predict the characteristics for variable witched reluctance motor.

The authors in [10] propose an alternative approach for variable reluctance motor is based on the decomposition of flux linkage into vector functions of rotor position and current.

This paper presents a novel analytical modeling for LVRM based on the flux linkage expression, taking account the non-linearity of the magnetic circuit.

[*]Aymen Lachheb, aymen.lachheb@enicarthage.rnu.tn

Then, the flux linkage expression is described by a Fourier series in which the three first components are considered, the coefficients of the terms in the Fourier series are determined by the characteristic of the flux as function of current obtained from the finite element analysis at the aligned, unaligned and middle position. Then, the effectiveness of the developed model is proved by 2D FEM results.

2. The structure and principal of LVRM

The motor investigated in this study is a three phases linear variable reluctance motor. A cross section of the machine is presented in Figure 1, which show the overall structure of the motor. It has a two salient construction parts the stator and the translator. The coils are concentrated around the poles of the stator and the translator has no windings, the three phase windings are designed by A, B and C. Therefore, each phase of this motor is composed of two poles and each pole has two coils in series.

Firstly, we consider that the poles of the phase A are aligned with the stator teeth, when current passed through the winding of phase B , a pull force is produced that tends to align the translator with stator poles which corresponds to a position with minimum reluctance.

In order to ensure a regular movement, each phase winding is energized by a available current at a suitable translator position. It means that the phases excitation is done by order from phase to the next phase as the translator moves from a shifted position to an aligned position.

For continuous motion the excitation sequence is A, B, C then A, B, C. Similarly, motion in the opposite direction can be produced by using the excitation sequence A, C, B, A, C, B.

2.1. Electric model for LVRM

For a variable reluctance actuator, the electrical model refers to a phase winding, can be described by a voltage equation.

Based on the Faraday's and Ohm's laws, the voltage supply the phase winding is given as:

$$U = RI + \frac{d\phi(x,I)}{dt} \tag{1}$$

Where, I is the current passing through a phase coils, R is the resistance per phase, x the position of translator and ϕ is the flux linkage of a phase.

2.2. Force equation

The force developed by a LVRM is relative to the rate of varaition of co-energy as long as the translator changes its position from one position to another. It can be defined as:

$$F_e(I,x) = \frac{\partial W_c(x,I)}{\partial x} \tag{2}$$

Where F_e is the instantaneous electromagnetic force and W_c is the co-energy. The latter can be expressed in terms of the flux linkage, as follows.

$$W_c = \int_0^I \phi(x,I)dI \tag{3}$$

The expression of the co-energy can be substitute in equation (2), which yields:

$$F_e(I,x) = \frac{\partial w_c}{\partial x} = \int_0^I \frac{\partial \phi(x,I)}{\partial x} dI \tag{4}$$

2.3. Mechanical equation

Before describing in detail the developed model, it is useful to present the equation governing the mechanical behavior of the linear motor.

The mechanical equation governing the linear motor operation can be given as:

$$m\,a = F_e - F_l - \xi\frac{dx}{dt} - f_0 \sin g(v) \tag{5}$$

Where m is the mass of the mobile, a is the acceleration, F_e is the developed force, F_l is the load force and ξ is the friction coefficient [11].

3. Finite Element Design of LVRM

The developed finite element model is performed according to the geometric parameters of the motor presented in table I.

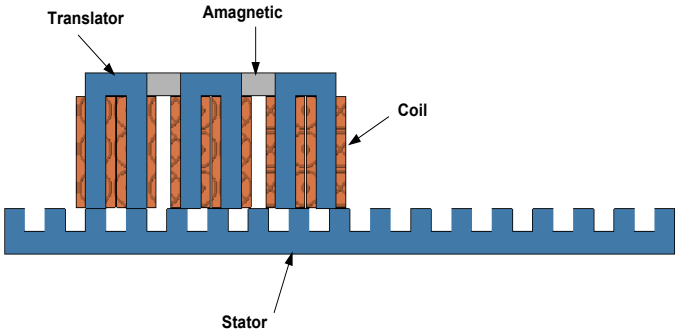


Figure 1. Single sided variable reluctance motor

To achieve a regular operation in step mode, the teeth and the slot widths are chosen equal.

$$\text{Hence } W_{tp} = W_{ts} = 30\text{mm} \tag{6}$$

The pole and slot widths are related to the pole pitch by the following equation:

$$\lambda = W_{ts} + W_{tp} \tag{7}$$

Where W_{ts} is the translator slot width, W_{tp} is the translator pole width and λ is the pole pitch.

The operation of the variable reluctance motor is based on the principle that the phases are powered independently and only one phase pole must be aligned with the stator poles. Then, to impose a regular step during the operation, a non-magnetic separation between different phases is necessary.

The translator is composed by three similar module separated by a non-magnetic material whose width is determinate by :

$$S = S_0 + k\lambda \tag{8}$$

With $0 < S_0 < \delta$ and $k = \{0,1,2,3\}$

The mechanical step δ_m is correlated to the pole pitch and the phases number N by the following equation (9):

$$\delta_m = |S_0 - W_{ts}| = \frac{\lambda}{N} \quad (9)$$

The yoke thickness of the stator and the translator are choosing equal to the slot width.

$$Y_{sy} = Y_{ty} = W_{ts} \quad (10)$$

Table 1: Specific geometry of designed motor

Parameters	Value(mm)
Stator pole width	30
Stator slots width	30
Stator yocke thikness	30
Translator pole width	30
Translator slots width	30
Translator yocke thikness	30
Separation	50
No of turns of wdg/phase	350
Section of copper	0.5
Air gap length	0.5

The 2D finite elements results obtained by the developed model are used to elaborate a surface response of the thrust force and flux linkage according to the translator positions and different supplying current level Figure 2 and Figure 3.

The simulation results are performed during a mechanical cycle where the moving part translates from the initial aligned position (x=0) to the next aligned position (x=0.06m) going through the unaligned position (x=0.03m).

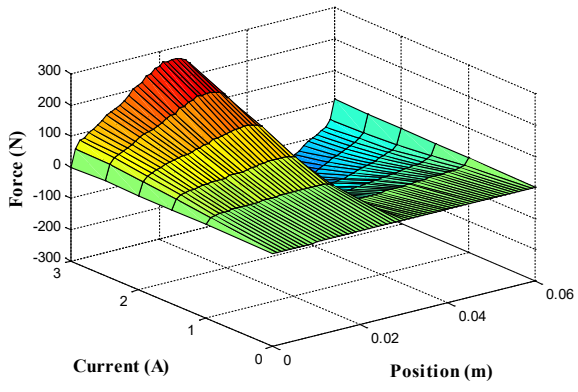


Figure 2 . Force as function of position for different current level

In order to develop a high force, the variable reluctance motor operates in saturation part of the B-H curve of the magnetic material, so it has a nonlinear behavior. The origin of this non linearity is mainly due to the non-linearity of the material used for the construction of magnetic circuits and the salient structure of the studied motor [12].

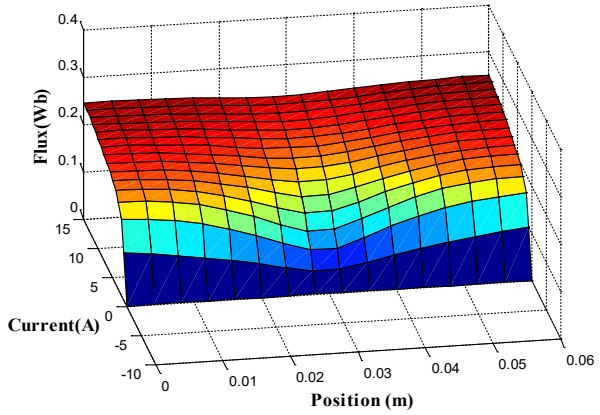


Figure 3 . Flux according to the translator position for different current

To avoid the problem of inaccuracies of the numerical method and in order to achieve a precise model of LVRM, it is recommended to use a methodology that permits to take account the nonlinearity of the magnetic characteristic of the used material while minimizing the simulation time. There are two main analytic modelling of LVRM used in previous work: the flux based model and the inductance based model [13, 14, 15].

In this study, the flux-based model is adopted to determine the static and the dynamic performances of the proposed motor. The proposed model for LVRM can be divided into two parts: a electrical and a mechanical part, they are realized as follows.

4. Analytical Modelling Approach for LVRM

In literature, there are two types of modelling, the linear model and the nonlinear model. Several research studies have adopted either the linear model [13] or the nonlinear model to predict the performance of a variable reluctance motor. The basic concept of this modeling consists of dividing the static flux linkage characteristics into two areas such as the unsaturated region and the saturated region as shown in Figure4. From the flux linkage characteristics, the two analytical models of the LVRM are defined as follows.

In the unsaturated region, when the phase current is lower than the saturation current $I_j \leq I_{sat}$, the LVRM is defined by a linear model. In the other hand, in the saturated region, when $I_j \geq I_{sat}$, the LVRM performances can be defined by a nonlinear model.

In the both models the flux linkage is represented by Fourier series and assumed has a cosine function of the position x although. However, in the nonlinear model a great harmonic order in Fourier series is used to give a satisfactory result. Then, it is possible to add another higher harmonics in the proposed model when more accuracy is needed [4].

4.1. Nonlinear Flux Based Model of LVRM

The modeling of variable reluctance machines is more complex than those of alternative current machines because of its highly nonlinear operation. The flux in each phase has a nonlinear

$2\pi/3$.

Consequently, an analytical model for the flux based on Fourier series representation is developed in this part.

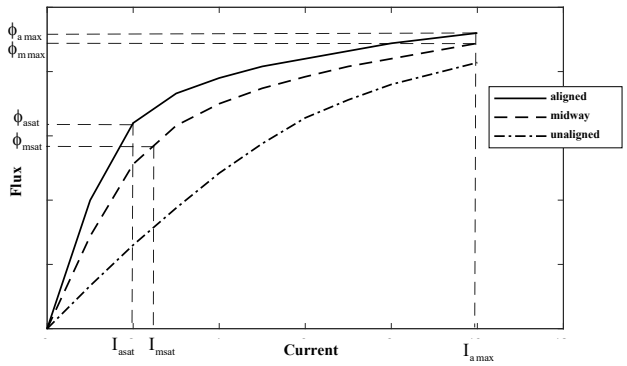


Figure 4. Flux current characteristic for three positions

Considering the first three components in Fourier series [7]. The flux in each phase can be represented as:

$$\begin{cases} \phi_j(x,I) = \sum_{k=0}^{n} \phi_k(I)\cos(k\frac{2\pi}{\lambda}x - (j-1)\frac{2\pi}{3}) \\ 0 < k < n \quad \text{and } j = 1,\ 2,\ 3 \end{cases} \tag{11}$$

Where n is the number of terms included in the Fourier series and j is the phase index.

Considering the first three components of the Fourier series in Equation (11), the flux expression become:

$$\begin{aligned} \phi_j(x,I) = \phi_0(I) + \phi_1(I)\cos(\frac{2\pi}{\lambda}x - (j-1)\frac{2\pi}{3}) \\ + \phi_2(I)\cos(\frac{4\pi}{\lambda}x - (j-1)\frac{2\pi}{3}) \end{aligned} \tag{12}$$

Where ϕ_0 , ϕ_1 , ϕ_2 are the first three terms of Fourier series.

The developed model is written in a matrix form as:

$$\phi_j(x,I) = \begin{bmatrix} \phi_0(I) & \phi_1(I) & \phi_2(I) \end{bmatrix} \begin{bmatrix} 1 \\ \cos(\frac{2\pi}{\lambda}x - (j-1)\frac{2\pi}{3}) \\ \cos(\frac{4\pi}{\lambda}x - (j-1)\frac{2\pi}{3}) \end{bmatrix} \tag{13}$$

The matrix model of flux is developed by considering three important positions, the aligned position (x=0), the midway position (x=15mm) and the unaligned position (x=30mm).

$$\begin{cases} \begin{bmatrix} \phi_{al} \\ \phi_m \\ \phi_{un} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \cos(\frac{2\pi}{\lambda}15) & \cos(\frac{4\pi}{\lambda}15) \\ 1 & \cos(\frac{2\pi}{\lambda}30) & \cos(\frac{2\pi}{\lambda}30) \end{bmatrix} \begin{bmatrix} \phi_0(I) \\ \phi_1(I) \\ \phi_2(I) \end{bmatrix} \\ \begin{bmatrix} \phi_{al} \\ \phi_m \\ \phi_{un} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \phi_0(I) \\ \phi_1(I) \\ \phi_2(I) \end{bmatrix} \end{cases} \tag{14}$$

With ϕ_{al} , ϕ_m and ϕ_{un} are the flux values for the aligned, intermediate and unaligned positions respectively.

Hence,

$$\begin{bmatrix} \phi_0(I) \\ \phi_1(I) \\ \phi_2(I) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \phi_{al} \\ \phi_m \\ \phi_{un} \end{bmatrix} \tag{15}$$

The three coefficients ϕ_0, ϕ_1, ϕ_2 can be computed based on the flux in three positions as follow.

$$\phi_0 = \frac{1}{2}\left[\frac{1}{2}(\phi_{al} + \phi_{un}) + \phi_m\right] \tag{16}$$

$$\phi_1 = \frac{1}{2}(\phi_{al} - \phi_{un}) \tag{17}$$

$$\phi_2 = \frac{1}{2}\left[\frac{1}{2}(\phi_{al} + \phi_{un}) - \phi_m\right] \tag{18}$$

The flux at the aligned and the midway position is determined by an arctangent function.

$$\phi_{al} = \frac{\tan^{-1}(\alpha_1 I)}{\alpha_2} \tag{19}$$

$$\phi_m = \frac{\tan^{-1}(\beta_1 I)}{\beta_2} \tag{20}$$

Where α_1, α_2 , β_1 and β_2 are constants that can be calculated by following the steps below:

- Define two points $\phi_{a\,max}$ and ϕ_{asat} on the flux curve relative to the aligned position.

The value of the flux $\phi_{a\,max}$ corresponds to the maximum current I_{max} (maximum current), whereas ϕ_{asat} it corresponds to the current I_{asat} (Saturation current).

-The constant α_1 can be determined by using the relation below:

$$\frac{\phi_{a\,max}}{\phi_{asat}} = \frac{\tan^{-1}(\alpha_1 I_{max})}{\tan^{-1}(\alpha_1 I_{asat})} \tag{21}$$

Where $\phi_{a\,max}$ is the maximum flux linkage at the aligned position, ϕ_{asat} is the saturated flux, I_{max} is the maximum current and I_{sat} is the saturated current.

-The constant α_2 can be determined as follow:

$$\alpha_2 = \frac{\phi_{sat}}{\tan^{-1}(\alpha_1 I_{sat})} \tag{22}$$

-The coefficient β_1 and β_2 can be calculated in the same way as α_1 and α_2 by using the suitable data obtained from the flux curve in the midway position.

-Since the saturation effect does not exist in unaligned position, therefore the flux linkage can be approximated by a linear equation.

$$\phi_{un} = L_{un} I \tag{23}$$

Where, L_{un} represent the phase inductance at the unaligned position.

4.2. Computation of Electromagnetic Force

The developed electromagnetic force is proportional to the square of current if the motor is unsaturated (L_a and L_{un}=cte), when the current increase and exceeds the saturation current then the electromagnetic force is no longer proportional to only the square of current it can be derived from the co-energy expression.

The flux depends to the translator position and the phase current. Then, the partial derivative of the flux with translator position can be written as:

$$\begin{aligned}\frac{\partial \phi_j}{\partial x} = &-\frac{2\pi}{\lambda}\phi_1 \sin\left[\frac{2\pi x}{\lambda} - (j-1)\frac{2\pi}{3}\right] \\ &-\frac{4\pi}{\lambda}\phi_2 \sin\left[\frac{4\pi}{\lambda} - (j-1)\frac{2\pi}{3}\right]\end{aligned} \tag{24}$$

From equation (4), the developed force by one phase of the LVRM is described as:

$$\begin{aligned}F_e &= \int_0^I \left[-\frac{2\pi}{\lambda}\phi_1 \sin(\frac{2\pi}{\lambda}x) - \frac{4\pi}{\lambda}\phi_2 \sin(\frac{4\pi}{\lambda}x)\right] dI \\ &= -\frac{2\pi}{\lambda}\sin(\frac{2\pi}{\lambda}x)\int_0^I \phi_1 dI - \frac{4\pi}{\lambda}\sin(\frac{4\pi}{\lambda}x)\int_0^I \phi_2 dI\end{aligned} \tag{25}$$

By substituting the equations (17) and (18) into equation (25), the expression of the force become.

$$\begin{aligned}F_e = &-\frac{2\pi}{\lambda}\sin(\frac{2\pi}{\lambda}x)\left[\frac{1}{2}\int_0^I \phi_{al} dI - \frac{1}{2}\int_0^I \phi_{un} dI\right] \\ &-\frac{4\pi}{\lambda}\sin(\frac{4\pi}{\lambda}x)\left[\frac{1}{2}\int_0^I \phi_{al} dI + \frac{1}{2}\int_0^I \phi_{un} dI - \int_0^I \phi_m dI\right]\end{aligned} \tag{26}$$

Where

$$\begin{cases}\int_0^I \phi_{al} dI = \int_0^I \frac{\tan^{-1}(\alpha_1 I)}{\alpha_2} dI = \frac{1}{\alpha_1\alpha_2}\left[(\alpha_1 I)\tan^{-1}(\alpha_1 I) - \frac{1}{2}\ln\left|1+(\alpha_1 I)^2\right|\right] \\ \int_0^I \phi_m dI = \int_0^I \frac{\tan^{-1}(\beta_1 I)}{\beta_2} dI = \frac{1}{\beta_1\beta_2}\left[(\beta_1 I)\tan^{-1}(\beta_1 I) - \frac{1}{2}\ln\left|1+(\beta_1 I)^2\right|\right] \\ \int_0^I \phi_{un} dI = \int_0^I L_{un} I dI = \frac{1}{2} L_{opp} I^2\end{cases} \tag{27}$$

5. LVRM Performances Prediction

To validate the model developed, we conducted simulation tests by developing a program designed around the Matlab/Simulink environment and adopting the following parameters:

U=24V, ξ =65Nsm^{-1}, $R = 8\Omega$, $\alpha_1 = 0.75$, $\alpha_2 = 6.55$, $\beta_1 = -0.54$, $\beta_2 = -6.59$, m=20Kg and L_{un}=0.5H.

In this section, the analytic model presented previously is simulated to predict the static and the dynamic performances of the LVRM. Subsequently, the obtained results are presented.

5.1. Static performances

The simulations results shown in this part are obtained basing on the analytical modeling approach described in part 3.

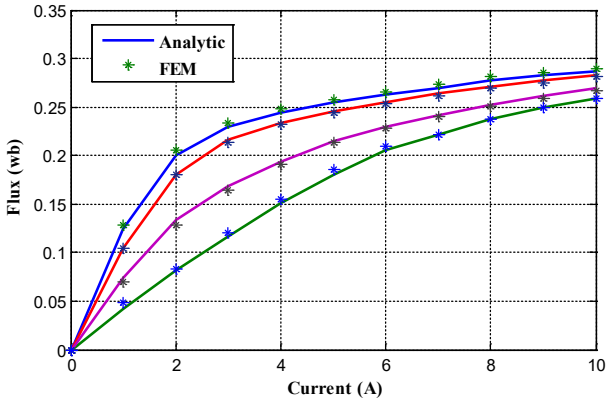


Figure 5 . Comparison between the calculated flux linkage and the FEM

Figure 5 shows a comparison of flux linkage/current characteristics for the designed motor of the analytic model with the result obtained by finite elements method (FEM). The comparison is made at four positions from aligned to completely unaligned position, it shows that the calculated flux is in good agreement with the computed FEM results.

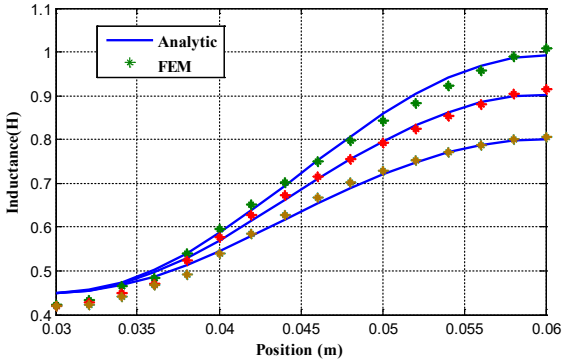


Figure 6 . Inductance as function of position for three current level

Figure 6 show the comparison of the computed magnetizing curves of the LVRM obtained by the proposed model and 2D Finite Element Method (FEM). Each curve of inductance is calculated under the condition that the phase A is excited by a selected current at different translator position. In addition, it is found that the obtained magnetizing curves from 2D FEM almost agree with those from the proposed model, only a slight difference in unaligned position is observed between the results obtained by both methods, due mainly to the assumption of non-existence of saturation at the unaligned position.

The phase inductance varies considerably at the aligned position with the current phase because of the magnetic saturation effect but the inductance at the unaligned position does not very much mainly because of the large reluctance that characterizes huge air gap in the flux path.

The accuracy of the proposed model can be enhanced by the appropriate choice of the number of Fourier terms. Thus, the observed mismatch can be improved by adding another Fourier term in the proposed model.

We can observe from Figure 6 that on the aligned position the values of the inductance vary considerably depending on the current level as the current increases the inductance decrease due to the saturation effect of the magnetic circuit and the inductance is constant at the unaligned position independent of the current value the effect of saturation is negligible.

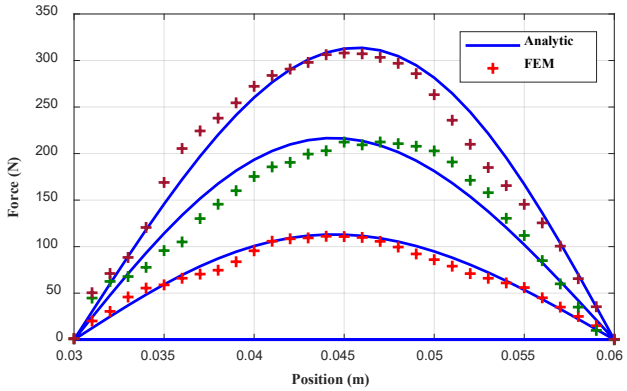


Figure 7 . Static force curve comparison for three current level

The comparative results of static force computed by the 2D FEM and calculated by an analytic model are shown in Figure 7, the comparison is done under the single phase operation at three different supplying currents from 1 A to 3A.

It can be observed from Figure 7, that the profile of the calculated force waveform relatively agrees well with that of the analytic one. Therefore, the validity of the proposed modelling approach is verified.

The developed model is validated from the above static behavior of the motor, so it will be possible to predict the dynamic performance of the motor by the proposed analytic modeling. In the following section, a non-linear dynamic performances analysis of LVRM are described.

5.2. Non-linear dynamic performances analysis :

In order to examine the validity of the proposed model in dynamic operating conditions, the analytical representation of flux is used to define the dynamic model by substituting this expression in the voltage equation. Then, the given model is implemented in Matlab Simulink to establish the dynamic performances of the LVRM and the obtained results has been compared with the FEM results. Therefore, a dynamic analysis based on the analytical model explained in this section.

For the variable reluctance motors, the flux linkage varies as a function of current and position, the phase voltage differential equation for active phase is given by.

$$U_j = RI_j + \frac{\partial \phi_j(x,I)}{\partial I}\frac{dI}{dt} + \frac{\partial \phi_j(x,I)}{\partial x}\frac{dx}{dt} \qquad (27)$$

Incorporating the flux expression (13) into equation (27), yields,

$$U_j = R_j I_j + \frac{\partial \phi_j}{\partial x} v + \left[\begin{array}{l} \frac{\partial \phi_0}{\partial I} + \frac{\partial \phi_1}{\partial I}\cos\left(\frac{2\pi x}{\lambda} - (j-1)\frac{2\pi}{3}\right) \\ + \frac{\partial \phi_2}{\partial I}\cos\left(\frac{2\pi x}{\lambda} - (j-1)\frac{2\pi}{3}\right) \end{array}\right]\frac{dI}{dt} \qquad (28)$$

Therefore, the derivative of current can be obtained by calculating the partial derivative of the flux with translator position and the partial derivative of the flux with excitation current as:

$$\frac{dI}{dt} = \frac{U_j - R_j I_j - \frac{\partial \phi_j}{\partial x} v}{\left[\begin{array}{l} \frac{\partial \phi_0}{\partial I} + \frac{\partial \phi_1}{\partial I}\cos\left(\frac{2\pi x}{\lambda} - (j-1)\frac{2\pi}{3}\right) \\ + \frac{\partial \phi_2}{\partial I}\cos\left(\frac{2\pi x}{\lambda} - (j-1)\frac{2\pi}{3}\right) \end{array}\right]} \qquad (29)$$

Where

$$\frac{\partial \phi_0}{\partial I} = \frac{1}{4}\left(\frac{\partial \phi_{al}}{\partial I} + \frac{\partial \phi_{un}}{\partial I}\right) + \frac{1}{2}\frac{\partial \phi_m}{\partial I} \qquad (30)$$

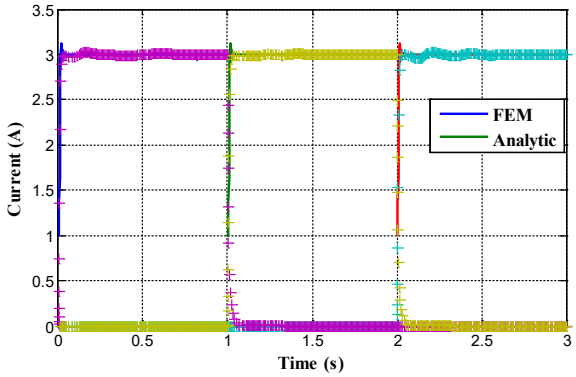


Figure 8 . Comparison of the computed and FEM phase current

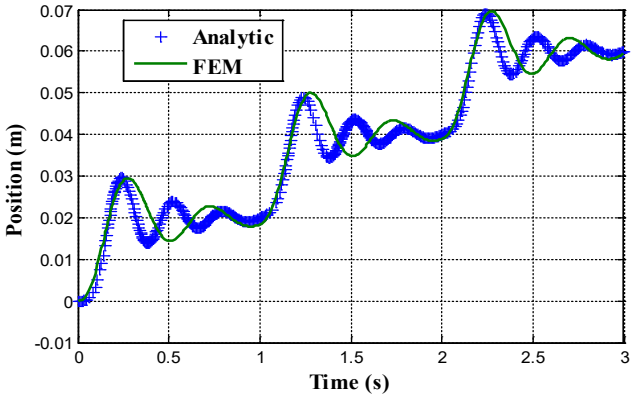


Figure 9 . Comparison of dynamic response of the motor

Figure 8 and Figure 9 present the comparison of the dynamic behavior of the motor. The superposition of these results shows that the dynamic response obtained by both methods admits the same response time with same over shoot, but slight difference noted for the number of oscillations between them. A rather good correlation in the results between the proposed model and FEM is observed.

6. Conclusion

In this paper, a novel non-linear model for a variable reluctance linear motor has been presented and compared with 2D FEM analysis results. This model based on the Fourier series analytical representation of flux linkage taking into consideration the effect of nonlinearity which has been neglected in some work describes its static characteristics. The model requires the motor flux characteristics obtained by finite element analysis.

The validity of the proposed model has been verified by the comparisons with results obtained from 2D FEM analysis. Furthermore, the accuracy of the proposed approach in dynamic conditions has been demonstrated. The developed model proves to be adapted for various uses with different machine configurations.

The developed model is motivated by the accurate performance prediction. In fact, it will be used to elaborate a control strategy for this motor.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] I .Mahmoud, H .Rehaoulia, "Design and modeling of open loop components for a biomedical application" International transactions On Electrical Energy Systems, 2016. https://**doi**.org/10.1002/etep.2204

[2] F .Daldaban, N .Ustkouncu, "A novel linear switched reluctance motor for railway transportation systems". Energy Conversion and Management,465-469,2009. http://dx.**doi**.org/10.1016/j.enconman.2009.10.009

[3] F R .Salmasi, B. Fahmi, "Modeling switched reluctance machines by decomposition of double magnetic saliencies". IEEE Transactions On Magnetics, Vol 40, No.3, 2004. https://ieeexplore.ieee.org/document/1298927

[4] A. Khalil, I. Hussain, "Fourier series generalized geometry based analytical model of switched reluctance machines". IEEE Transactions On Industry Applications, Vol 43, No.3, 2007.

[5] A .Lachheb, L .El Amraoui, J. Khediri, "Finite element modelling of linear motor for automatic sliding door application". International journal of advanced computer science and applications, Vol.7 No.8, 2016.

[6] B .Fahimi, G .Suresh, J .Mahdavi, M. Ehsani, "A new approach to model switched reluctance motor drive application to dynamic performance prediction control and design". IEEE Powers Electronics, 1998.

[7] D. A. Andrade, R. Krishnan, "Characterization of switched reluctance machines using fourier series approach". 36th Conference Industry Applications, USA,2001. https://ieeexplore.ieee.org/document/955391

[8] H .Gao, F.R .Salmasi, M. Ehsani "Inductance model based sensorless control of the switched reluctance motor drive at low speed". IEEE Transactions On Power Electronics, Vol.19, No.4, 2004.

[9] H.P .Chi, R.L .Lin, J. F. Chen, "Simplified flux linkage model for switched reluctance motors" IE Proc Electr Power Applications, 152 (3), 2005.

[10] P. Lobato, S .Rafael, J. Pires, "Flux linkage characteristics models for switched reluctance machines". IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society 2010.

[11] W. Zaafrane, J. Khedhiri, H. Rehaoulia, "Low-cost linear switched reluctance motor velocity and position control". Journal of Circuits Systems and Computers, 2016.

[12] M. Alrifai, M. Zribi, R. Krishnan , M .Rayan, "Nonlinear Speed Control of Switched Reluctance Motor Drives Taking into Account Mutual Inductance". Hindawi Journal of Control Science and Engineering, 2007.

[13] S. Buja, M. I. Valla, "Control characteristics of the SRM drives operation in the linear region". IEEE Transactions on Industrial Electronics, Vol.38 No.5, 1991.

[14] Z. Zhu, Y. Sun, Y. Yuan, "Decoupling control for fual winding Bearingless switched reluctance motor based on improved inverse system method". Hindawi Journal Mathematical Problems in Engineering, 2017.

[15] S.Husain, A.Hossain, "Modeling simulation and control of switched reluctance motor drives" IEEE Transactions On Industrial Electronics, Vol 52, No.6, 2005.

[16] S .Mir, I .Husain, M. E .Elbuluk, "Switched reluctance motor modeling with on-line parameter identification" IEEE Transactions On Industry Application, Vol.34, No.4, 1998.

[17] J .Fang, Y. Ren, "Decoupling control of magnetically suspended rotor system in control moment gyros based on an inverse system method". IEEE/ASME Transactions on Mechatronics 2012; Vol. 17, No. 6, pp. 1133–1144, 2012.

[18] Siadatan A, Torkaman H, Afjei E, "Septic segment switched reluctance machine design modeling and manufacturing". International Transactions on Electrical Energy Systems, 2016.

[19] J .Faiz, K M .Zadeh, "Design optimization of switched reluctance machines for starter generator of hybrid electric vehicle by genetic algorithm". European Transactions on electrical power 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 3, 194-197 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Digitization Impact Assessment Model for Secondary Schools: Case of Nairobi County in Kenya

Nicodemus Kalugho Mwambela, Simon Nyaga Mwendia*

KCA University, Faculty of Computing and Information Management, 00200, Kenya

ARTICLE INFO

Article history:
Received:15 December, 2018
Accepted:25 May, 2019
Online: 18 June, 2019

Keywords:
Digitization impact index
Secondary school
Model

ABSTRACT

Digitization is viewed as the capability to utilize digital technologies for generating processing, sharing and transacting information. Studies show that digitization has now become ubiquitous. In almost all the sectors, we routinely interact with digital technologies. As a result, the impact of digitization is being observed in such sectors. For instance, the positive impact of digitization in education sector includes more efficient administration, reduction in workload, better accessibility to information, better management of school resources and enhanced quality of reports. However, in the context of secondary schools, the literature is conspicuously silent on the formal way of assessing the impact of digitization on both the academic performance and service delivery. As an attempt to address the gap, this paper describes a model that can evaluate the impact of digitization in such schools. Python programming tools were used to demonstrate implementation of the model. The model can help governments to formulate policies related digitization of secondary schools and can also be used by funding organization to demonstrate accountability when making investment decisions.

1. Introduction

According to [1] digitization refers to the capability to utilize digital technologies for generating, processing, sharing and transacting information. Previous studies show that digitization in secondary schools is realized by integrating information communication technology(ICT) tools like School Management information systems (MIS) to enhance efficiency of providing administration services. Examples of such services are formulating strategic plans, evaluating staff performance, communication and distribution of resources [1, 2].

Integration of ICT tools in education calls for the development of indicators to monitor the impact of such tools. Both funding sources and the public can use the indicators assess accountability. In addition, there is a need for tools for indicating how use of technology in education is promoting empowerment, creativity, equality, efficiency among learners. Many academic researchers have attempted to address these needs at theoretical and empirical levels but have encountered two difficulties. First, student performance is hard to measure and still there is no common understanding about its definition. Second, ICT is characterized by changing or evolving technologies and their impacts are difficult to separate from their setting. Therefore, the association between the use of ICT and educational performance is not clear, and inconsistent findings are presented in the literature [3]. To address some of these difficulties our study aims at establishing an appropriate model for assessing the digitization level and its impact on education institutions including secondary schools.

* Simon Nyaga Mwendia, Email: smwendia@kca.ac.ke

2. Related Work

2.1. Digitization

Education is one of the important mechanisms that provide the necessary skills, competencies and knowledge among people. Consequently, research on factors influencing the academic performance of students in education institutions has become a topic of increasing interest. Previous studies have reported that digitization and the integration of ICT tools is one of the factors that are attributed to improvement of academic performance in secondary schools [5]. A recent report from the Education Council of the Netherlands has revealed that the education sector is still looking for the suitable level and appropriate methods of digitization [6]. This observation has created the need for establishing an appropriate index for measuring digitization level and its impact on education. Examples of existing models of indices include digitization index [7] and digitization impact index.

a) Digitization Index (DiGiX)

According to [7], Digitization Index (DiGiX) can be described as a composite index for measuring the impact of digitization in societies such as learning institutions. This is a tool assessing the extent to which ICT is integrated in institutions[8]. Previous research work has been focused on developing indices that can assess digitization in a country. An example is Digitization Index proposed by [9] as tool for assessing the level digitization. This level can be measured using six main attributes. That is, (i) Ubiquity, which refers to the level of accessing universally digital applications and services by consumers and enterprises, (ii) affordability, which is described as the degree to which digital services are valued in a range that makes them available to a large population, (iii) reliability that is viewed as the quality of available digital services, (iv) speed, which is described as the extent to the degree to which digital services can be available in real time, (v) usability that is described as the simplicity of using digital services and the capacity of local ecosystems to improve utilization of these services and, (vi) skill that is viewed as the ability of users to incorporate digital services into their daily lives and businesses. Figure 1, illustrates how each of the metrics are related in digitization index.

c) Digitization Impact Index

Digitization Impact Index can be described as the effect of digitization of a society such as a school or a country. In the context of a school, Digitization Impact index can be studied as an impact of digitization of a school in a user satisfaction model. In this case, the user is the staff or student who is directly impacted by the utilization of digital services in the school. Key indicators to compute their level of satisfaction are derived from the rating of the efficiency of services offered in the school which would be influenced by the level of digitization. According to [10] people are used to the concept of rating things with numerical scores and these can work well in surveys. Once the respondent has been given the anchors of the scale, they can readily give a number to express their level of satisfaction. Typically, scales of 1- 5 are used where the lowest figure indicates extreme dissatisfaction and the highest shows extreme satisfaction.

Several studies have observed that digitizing schools has an impact on education that can be measured using the following indicators:

i. Improvement of students' academic performance that is generally evaluated by the mark or grade attained by the student in an examination [5]
ii. Increase of enthusiasm, interest and creativity [5].
iii. Ease of system use like explaining concepts when cast on screen [5].
iv. The speed of information access increase as a result of internet infrastructure that support pervasive, simultaneous utilization of devices for instruction, evaluation, and school operations [12].
v. The number of decisions to make increases. These include device acquisition, content options, the role of technology in supporting classroom management, security and privacy, and data management [12].

However, policy makers are confronted with two main challenges that are related to application of digital index. First, there are no standard performance indicators to measure the level at which ICT is incorporated in societies. The second challenge is that there is lack of instruments for measuring the impact of adopting mass connected digital technologies and applications in societies and economies [8]. Consequently, there is lack of a standard index for assessing the impact of digitization in secondary schools.

3. Proposed Model

As an attempt to address these challenges, we propose a model that can be used to assess the level and impact of digitization in a typical secondary school. The model was derived by combining a composite digitization index proposed by [9] with indicators of digitization impact stated by [5, 12]. This is shown in Figure1.

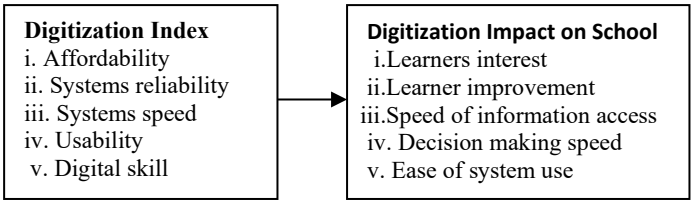


Figure1: Digitization impact conceptual model

As illustrated in Figure1, *digitization index* is representing independent variable for describing the level of digitization in a typical a secondary school. On the other hand, *digitization impact on the school* index represents the dependent variable. The proposed model can also be illustrated using simple regression equation of the form, $\mathbf{Y= a + b_1X_1}$, where Y is the *digitization impact index* on the school , a is the constant parameter, b_1 is the parameter to be estimated and variable X_1 is the *digitization index,* an independent variable.

4. Significance of the Study

There are three main benefits that are associated with the proposed model. First, it can be used by the government to formulate policies related digitization of secondary schools, particularly in the African contexts like Kenya. The second benefit is that it can be used by funding organization to demonstrate accountability when making investment decisions [3]. The third benefit is that it can be used by school managers to assess the impact of integrating ICTs in a typical secondary school.

5. Study Method

According to [13] as cited in [14], at least thirty (30) subjects required for correlational research to investigate relationship between variables and at least 10% of the accessible population is required for descriptive research. Based on this argument, the researcher collected data from 35 teachers in 11 (>50%) schools of Makadara Sub- County in Kenya.

The study sought the teachers input in assessing the reliability and speed of computer systems, adequacy of infrastructure and the impact that this had on both their students' interest and improved academic performance. The input from teachers was solicited through questionnaires and converted into numeric values by use of Likert scale.

For each of the seven (7) schools visited, individual responses obtained through questionnaires were used to generate both Digitization index and Digitization impact index. This was done by computing the percentage of arithmetic mean obtained from 5-point Likert scale. Based on the scale, the maximum rating for each indicator was five (5) and the lowest rating was one (1). Therefore, the maximum possible sum of the five identified indicators was 5X5=25. The sum of the levels was then expressed as a percentage to the maximum possible sum (25) to obtain an index for each school. For instance, if an individual respondent rated Learner interest = 3, Learner improvement =3, Speed of information access = 4, Ease of system use =3 and Decision making speed =2, then the ICT Integration index can be computed as follows:

$$X = \frac{(3+3+4+3+2)}{5X5} \times 100\ \% = \frac{15X100\%}{25} = 60$$

During data analysis, digitization index was treated as independent variable while academic performance was treated as a dependent variable. The internal consistency of the questionnaire was tested by use of Cronbach Alpha measure and found to have the acceptable value of 0.735. Correlation analysis was then carried out to assess the strength and the direction of association between variables [15] while linear regression analysis were used to develop the proposed model.

6. Results

Table1 presents shows results from the seven (7) schools, where variables X_1, represents Country's' Digitization index, and Y represents digitization impact on learners behavior.

Table1: Digitization index (X_1) and Digitization Impact index (Y)

Digitization index (X_1)	Digitization Impact index (Y)
64	44
80	80
80	76
72	92
92	60
68	72
52	52

Pearson correlation results from the seven (7) schools' sample showed that Digitization index was positively related to digitization impact with a correlation coefficient of 0.38. This is shown in Figure 2.

	Digitization_Index	Digitization_impact
Digitization_Index	1.000000	0.379779
Digitization_impact	0.379779	1.000000

Figure 2: Correlation between digitization index and digitization impact

Findings from Sklearn library that is implemented in python programming language revealed that academic slope coefficient (b_1) was approximately 0.49 while the intercept coefficient was 32.20. The two computed coefficients were then substituted in the regression equation to formulate the empirical model that is shown in Figure 3. The model means that for every 0.49 increase in the digitization index there is one unit increase in the academic slope.

$$Y = 32 + 0.49X_1$$

Figure 3: The developed empirical model

For the purpose of enhancing understanding Seaborn and Matplotlib packages were used to visualize the developed model from the seven schools sample. The results are shown as shown in Figure 4.

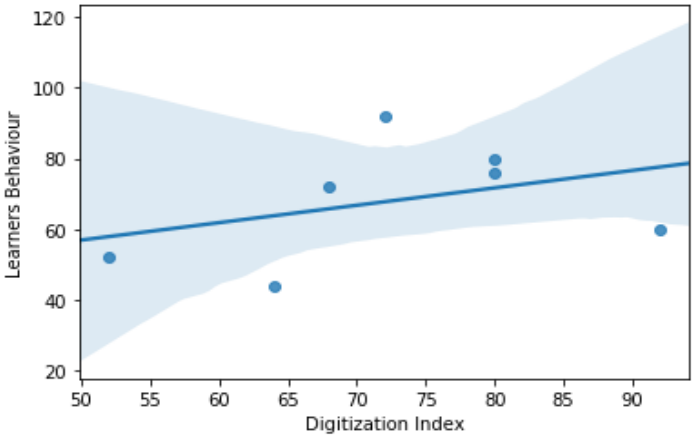


Figure 4: Visualization of the developed empirical model

7. Conclusion and Future Work

Findings from this study reveal that Digitization index of secondary schools can be measured using five indicators. That is, (i) affordability, (ii) systems reliability, (iii) systems speed, (iv) usability and (v) digital skill. In addition, linear regression equation can be used as an index for measuring the impact of digitization on both the academic performance and service delivery. Metrics for measuring such an impact include (i) Learners interest, (ii) learner improvement, (iii) speed of information access, (iv) decision making speed, and (v) ease of system use

From the reviewed literature it was observed that digitization has a positive impact on academic performance, interest and enthusiasm, ease of system use, and decision making process [5, 12]. On the other hand, the evaluation results of the proposed model showed that there is positive correlation between digitization and academic the effect on same indicators. Therefore, it can be concluded that the proposed model is consistent with findings of [5, 12].

Further research can be conducted to explore other measures for assessing digitization impact. The generalization of the proposed model can be evaluated by testing it in other education contexts like colleges and universities.

Conflict of interest

The authors declare that there is no conflict of interests in relation to the publication of this paper.

References

[1] R. Bisaso and A. Visscher, "Computerised School Information Systems Usage in an Emerging Country - Uganda," in *Information and Educational Management in the Knowledge Society*, New York,USA: Springer, 2005, pp. 81–98.

[2] M. Shah, "Impact of management information systems (MIS) on school administration: What the literature says," in *5th World Conference on Educational Sciences - WCES 2013*, Rome,Italy, 2013, pp. 2799 – 2804.

[3] A. B. Youssef and M. Dahmani, "The Impact of ICT on Student Performance in Higher Education: Direct Effects, Indirect Effects and Organisational Change," *Univ. Knowl. Soc. J. RUSC*,vol.5,no1, 2008.

[4] R. Katz, P. Koutroumpis, and F. M. Callorda, "Using a digitization index to measure the economic and social impact of digital agendas," *emeraldinsight*, vol. 6, no. 1, pp. 32–44, 2014.

[5] F. K. Kisirkoi, "Integration of ICT in Education in a Secondary School in Kenya: A Case Study," *Lit. Inf. Comput. Educ. J. LICEJ*, vol. 6, no. 2, 2015.

[6] VSNU, "Digitisation in academic education." VSNU Association of Universities,the Netherlands, 2017.

[7] N. Camara and D. Tuesta, "DiGiX: the Digitization Index." BBVA Research, 2016.

[8] R. Katz, K. Pantelis, and M. C. Fernando, "Using a digitization index to measure the economic and social impact of digital agendas," *Emeraldinsight*, vol. 16, no. 1, pp. 32–44, 2014.

[9] K. Sabbagh, B. El-Darwiche, R. Friedrich, and M. Singh, "Maximizing the impact of digitization," Booz&Company, 2012.

[10] P. Hague and N. Hague, "Customer Satisfaction Surveys & Research: How to Measure CSAT." B2B International, n.d.

[11] N. Bitner and J. Bitner, "Integrating Technology into the Classroom: Eight Keys to Success," *Jl Technol. Teach. Educ.*, vol. 10, no. 1, pp. 95–100, 2002.

[12] K. Khemani, C. Hagen, B. Ross, and A. K. Jamjourm, "Digital School." ATKearney, 2013.

[13] L. . Gay and P. . Diehl, *Research Methods for Business and Management*. Macmillan, 1992.

[14] R. Hill, "What Sample Size is 'Enough' in Internet Survey Research?," vol. 6, no. 3–4, 1998.

[15] O. Mugenda and A. Mugenda, *Research methods. Quantitative and Qualitative approaches. Nairobi.* Nairobi: Africa Centre for technology studies, 2003.

.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 3, 198-206 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Indoor Positioning and Path Planning Platform for iRobot Create 2 Sweeping Robot

Rung-Shiang Cheng[*,1], Wei-Jun Hong[3], Jhe-Lin Li[2], Kawuu W. Lin[3]

[1]*Department of Information Technology, Overseas Chinese University, Taiwan, R.O.C*

[2]*Department of Computer and Communication, Kun Shan University, Taiwan, R.O.C*

[3]*Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, R.O.C*

ARTICLE INFO

Article history:
Received:19 December, 2018
Accepted:25 May, 2019
Online: 18 June, 2019

Keywords:
Indoor position
Path control
Sweeping robot

ABSTRACT

Many robots have been proposed for smart applications in recent years However, while these robots have good ability to move between fixed points separated by short distances, they perform less well when required to navigate complex interior spaces with an irregular layout and multiple obstacles due to their relatively crude positioning and path control capabilities. Accordingly, the present study proposes an integrated system consisting of a Raspberry Pi development platform, various sensor devices and iBeacon technology to facilitate the path control, indoor positioning, and obstacle avoidance of a programmable iRobot Create 2 sweeping robot. The capabilities of the proposed system and its various components are investigated by means of practical experiments and numerical simulations. In general, the results confirm that the proposed integrated system provides a viable platform for the future development of sophisticated indoor robots for smart indoor applications.

1. Introduction

In recent years, robot technology has been introduced into related application of smart network. However, interior space is a very complicated environment because signal transmission can be easily influenced by architectural structure. Thus, for many years, well-known services to the mess such as GPS positioning system or Google map are most applied in outdoor public infrastructure instead of interior space. This has blocked the application of robots in smart homes; at present, smart robots mostly move between fixed points, on certain tracks or only in short distance; for the movement in interior space of robots, there have not proper solutions because researches in this field is still in startup stage and robots are in lack of the support in IoT environment; plus, problems emerged when it comes to the sensing technology, spatial orientation, path planning, obstacles avoidance, robot learning and communication between robots still need to be solved and overcome by further integration.

To introduce the application of robots from factories to homes, we need to enable smart robots to arrive the right position rapidly through the indoor application service or precisely identify robots or users via indoor positioning navigation. Thus, it is necessary to conduct further research on smart robots to develop a positioning service system used indoor. In this way, smart robots can be extended from industrial application in factories to caring service in smart homes, bringing more diverse possibilities to home application service of IoT.

This study is organized as below: Section 2 introduces related work on indoor positioning, path planning and searching, avoiding obstacles and automatic following; Section 3 introduces research steps and system implementation; Section 4 represents conclusions.

2. Research on related application of home robots

How to let robots assist people to deal with daily tasks in various scenarios and precisely identify their location and arrive target areas to complete tasks will be important goals to introduce robots to smart home application. The following are some related study of home robots.

2.1. Indoor positioning

In view of this, [1] proposed to utilize RSSI signal of wireless network and triangulation algorithm as be indoor positioning

[*]Corresponding Author: Rung-Shiang Cheng, rscheng@ocu.edu.tw

mechanism for robots. The results showed the positioning precision is 0.125 m and the error is below 0.4%.

In [2], the author proposed to utilize laser distance meter to construct indoor environment model. The meter can measure its moving path and calculate via the assistance of cloud server, effectively avoiding obstacles on moving path. In [3], it proposed to utilized the information change in the sensor or the robot and calculate its moving direction and location. Through change values measured by gyro and accelerometer, the robot can calculate its X, Y and θ. After calibration, it can obtain the reference coordinate of robots and plan an effective moving path.

2.2. Path planning

Coverage path planning [4] algorithm for robots is an important efficiency indicator for home robots; random path planning is an intuitive way to move of robots. Robots forward linearly towards any direction and change path only when hitting on obstacles and repeat this process until it reaches maximum coverage scope. Such algorithms have the advantage of simplicity since very few sensors are required and there is no need to construct and maintain an indoor environment map [5]. However, they require a long time to achieve full coverage of the target environment; DmaxCoverage [6] an algorithm which can assist robots to construct environment map to reach maximum coverage scope; furthermore, in [7], this study utilized Dijkstra's algorithm to generate optimum path in environment to avoid the decrease of efficiency during random path planning.

2.3. Search, obstacle avoidance and automatic follow-up

Unlike robots, humans can use their own sense organs to create space and sense of distance. Therefore, in [8] this paper implements a mobile robot with an automatic following system. The mobile robot can perform positioning, mapping, and moving the target side. Tasks such as distance, human detection and target tracking, and can be applied in real life, such as museum, office or library guidance; in [9], this paper implements a wheeled robot that can perform search and obstacle avoidance work. The robot is equipped with a vision and distance sensor, which can match multiple objects by visual system identification objects to create multiple hypothesis blocks for the environment. These multiple hypothesis blocks can make the robot avoid obstacles and explore the target object.

3. System design and implementation

Figure 1 shows the functional block diagram of the implemented robot. The function of positioning is a crucial factor to introduce robots into interior application. To achieve this goal, this study combines Raspberry Pi with BLE Bluetooth communication technology to implement iBeacon of robots; when the robot moves, if it detects iBeacon and reports the positioning information, the system is able to know the location of the robot.

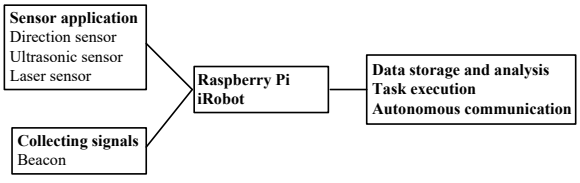


Figure 1: System block diagram of the sweeping robot

Figure 2 shows the original look of programmable sweeping robot. This study combines iRobot Create 2 programmable sweeping robot with Raspberry Pi programmable control electric circuit with I/O device to explore straight movement and rotation of robots. We also use Python to write tests of every control application. developing a smart robot which is equipped with indoor positioning function; Figure 3 is the prototype of the robot for indoor application.

Figure 2: Programmable Sweeping Robot (Original Look)

Figure 3: Smart Robot Developed by Combining iRobot Create 2 Programmable Sweeping Robot with Raspberry Pi

3.1. Composite sensing device

Linear walking and rotation are the basic movement instructions of the robot. In order to understand the difference between the expected result and the actual result after the instruction is executed, this study installed sensing devices such as distance, magnetic field, acceleration, and gyroscope on the cleaning robot. This study first measure the robot's motion data and then use the analysis results to correct the robot's motion parameters.

Show as Figure 4, take the orientation sensor as an example. After the robot adds the orientation sensor, the system developer can use the sensor's numerical change to determine the robot's movement state; because the X axis represents the robot's direction of travel, the robot can use it to determine the robot when performing a task. The forward direction; on the other hand, since the Z axis of the accelerometer is negative when the robot's operation panel is upward and downward when it is positive, it is possible to use the Z axis of the acceleration sensor to detect whether the robot is running when the robot is operated. Overturned or tilted state.

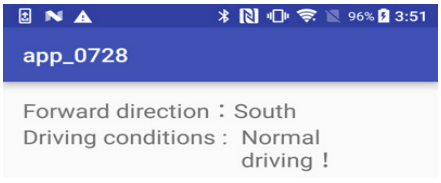


Figure 4: The result of the smart robot orientation and the tilt state of the car body

3.2. Robot moving path planning mechanism

To enhance the accuracy of control command, we need to measure reaction delay time and error when executing assigned process after robots receive commands.

3.2.1. Measure and Observe the Straight Movement

To enhance the accuracy of control command, we need to measure reaction delay time and error when executing assigned process after robots receive commands.

Assume the program execution time is 10.05 seconds to 11.00 seconds, and forwarding speed of the robot is set as 200 mm/sec; Table 1 shows the measured moving distance of the robot at different second.

As shown in Table 1, the experimental results indicate after the control device gives commands to the robot, the actual complete time of robots executing commands will have delay for about 0.5155 second (time error) and the average value of distance error is 10.1662 cm. This result can be used by the system to modify parameters so as to enhance the accuracy of positioning of the robot.

Table 1: Computed Measured Results of Experimented Robot Move Straightly for 11 Seconds

Execution Time	**Time Error**	**Computed Distance**	**Actual Distance**	**Distance Error**
10.5657	0.5157	211.3148	200.0	11.3148
10.6153	0.5153	212.3066	201.9	10.4066
10.7120	0.5120	214.2402	203.5	10.74024
10.8067	0.5067	216.1356	205.9	10.2356
10.9189	0.5189	218.3784	208.0	10.3784
11.0167	0.5167	220.3347	210.0	10.3347
11.1192	0.5192	222.3845	212.0	10.3845
11.2163	0.5163	224.3278	214.3	10.0278
11.3168	0.5168	226.3371	217.0	9.3371
11.4165	0.5165	228.3313	219.0	9.3313
11.5168	0.5168	230.3377	221.0	9.3377
	0.5155			**10.1662**

According to the above measurement it is found when speed is fixed, we can obtain stable second difference and stable distance difference. Thus, when controlling the moving distance of robots, we can do corrections by substituting values into the Eq.(1):

$$estimated\ time = \left(\frac{distance+difference}{speed\times 0.1}\right) - offset\ time \quad (1)$$

Amongst, distance indicates predetermined moving distance, difference indicates error correction distance, offset time indicates delay time of actual execution; from measured results in Table 2 it is known: let difference = 10.1662, offset time = 0.5155, assume predetermined straight moving distance of the robot is 800 cm and then the values adding into the Eq.(2):

$$estimated\ time = \left(\frac{80+10.1662}{200\times 0.1}\right) - 0.5155 = 39.9928 \quad (2)$$

Through this equation we obtain predetermined execution time of the program is 39.9928 seconds; after the program is execute, we obtain the actual execution time of the robot 40.5387 seconds, moving distance is 799 cm; the measured results show after adding it to the correction equation, the actual moving distance of the robot 799 cm is close to the predetermined value 800 cm, and difference is about 0.125%.

Table 2: Measured Results after Adding the Straight Walk to the Correction Equation

Actual Seconds	Second Difference	Actual Distance
40.5387	0.5459	799

3.2.2. Measure and Observe the Rotation Angle

To observe whether the rotation angle and direction of the robot are as what we have predicted, we combine the mobile sensor with the robot. Through function of iPhone DeviceMotion we obtain the measurement data of the robot; experimental data here will be compared with that obtained from SenseHat compound sensor; Table 3 shows the measured results by observing rotation angle per second after the robot finishes straight move and another move for 1.1~3.3 seconds at 100 mm/sec rotation speed. The testing process is shown in Figure 5.

(a) The Status of Actual Measurement

Device Motion		Magnetometer	
Pitch:	1	X:	26.18
Roll:	-2	Y:	-124.12
Yaw:	-35	Z:	-43.10
Accelemeter		**Gyroscope**	
X:	-0.006	X:	-0.049
Y:	-0.035	Y:	-0.085
Z:	-0.970	Z:	-0.029

(b) Measured Results

Figure 5: Measure Rotation Angles of Robots

From the experimental results we know when rotation speed is fixed, dividing actual angle by program seconds will obtain rotation angle per second, rotation angle per second is stable and thus we obtain the equation below:

$$excution\ time = expected \div average \quad (3)$$

Let expected denote the predetermined rotation angle, average denote the actual angle/ obtained average rotation angle according to program execution seconds (control factor) According to the measured results in Table 3, let average = 55.21061296□ /sec; during the execution, we simply need to substituting the degree of

angle to rotate into the above-mentioned equation to obtain the required seconds for rotating to the assigned angle; taking rotating 90-degree angle as example, through the equation we obtain the execution time of the program will be expressed as execution time = 90 / 55.21061296 = 1.6301 seconds.

Table 3:: Actual Measured Results of Rotation Angles of Robots

Execution Time in Seconds	Rotation Angle	Rotation Angle Per Second
1.502199	55	55
1.602238	61	55.45454545
1.702244	67	55.83333333
1.802489	72	55.38461538
1.902531	77	55
2.002577	83	55.33333333
2.102841	88	55
2.202954	93	54.70588235
2.302842	100	55.55555556
2.403040	104	54.73684211
2.503230	109	54.5
2.603253	117	55.71428571
2.703390	122	55.45454545
2.803547	126	54.7826087
2.903717	132	55
3.003546	139	55.6
3.103666	142	54.61538462
3.203762	149	55.18518519
3.303927	153	54.64285714
3.404050	161	55.51724138
3.504325	165	55
3.604233	171	55.16129032
3.704476	176	55
		55.21061296

Table 4 shows the results obtained when the robot rotated 90 degrees, 180 degrees, 270 degrees, and 360 degrees. It can be seen from the experimental results that with the increase of the program execution time, the errors generated by the robot after the execution of the rotation action will gradually appear.

Table 4: Actual results of the angles obtained after actual implementation

Estimated rotation angle	Calculated time	Actual execution time	Actual rotation angle	Angle error
90	1.6301	2.132857	91	1
		2.133154	90	0
		2.132777	91	1
		2.133019	87	3
		2.132970	91	1
		2.132866	90	0

(a) 90 degrees

Estimated rotation angle	Calculated time	Actual execution time	Actual rotation angle	Angle error
180	3.2602	3.762850	180	0
		3.764159	183	3
		3.762853	180	0
		3.763769	181	1
		3.764353	181	1
		3.764293	180	0

(b) 180 degrees

Estimated rotation angle	Calculated time	Actual execution time	Actual rotation angle	Angle error
270	4.8904	5.396417	269	1
		5.396603	272	2
		5.396394	268	2
		5.396588	268	2
		5.396419	271	1
		5.396533	270	0

(c) 270 degrees

Estimated rotation angle	Calculated time	Actual execution time	Actual rotation angle	Angle error
360	6.5205	7.028154	357	3
		7.028164	362	2
		7.028282	362	2
		7.028401	362	2
		7.028339	362	2
		7.028357	362	2

(d) 360 degrees

SenseHAT is an expansion board for the Raspberry Pi. As shown in Figure 6, since the robot's magnetic field will affect the SenseHAT's value, in order to improve the accuracy, SenseHAT will calibrate the magnetic field, and then measure the SenseHAT detection value is correct. Table 5 uses the rotation angle of 72 degrees as an example to judge the difference between the predicted angle and the actual angle. The maximum value of the experimental result drops to 5, so it needs to be corrected after each rotation command

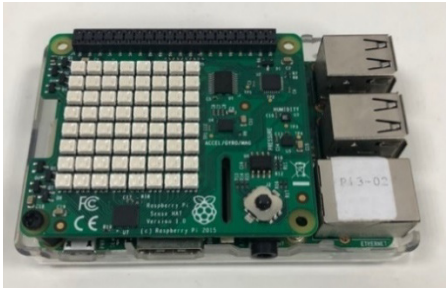


Figure 6: SenseHAT composite sensor

3.2.3. *Measure and Observe the Rotation Angle*

As mentioned above, the movement of robots is consisted of two parts including straight movement and rotation. Thus, we can write its moving path into script and then through changing

commands to make robots execute tasks. Let s denote straight moving, t denote back moving, r denote right turn and l denote left turn. Taking Figure 7 as example, the script of the moving path of the robot will be written into "s290 r90 s47 l90 s310 r30 s122." if we break down the script by rules, it can be divided into seven command list including forwarding 290 cm, rotating right for 90-degree angle, forwarding 47 cm, rotating left for 90-degree angle, forwarding 310 cm, rotating right for 30-degree angle, forwarding 122 cm and arriving the destination.

Table 5: Experiment with a rotation of 72 degrees

Initial angle	Estimated rotation angle	Actual rotation angle	Angle error
105	177	180	3
	249	251	2
	321	323	2
	33	38	5
	105	106	1
	178	181	3
	250	253	2
	322	323	1
	34	38	4
	106	106	0

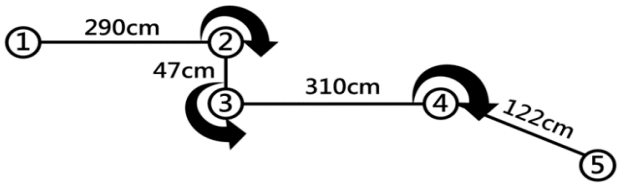


Figure 7: Moving Path Planning of Robots

3.2.4. *Production of script tools*

Although the path of the robot can use the previously written script to write the relevant parameters to the robot. However, in order to make the robot's control more flexible, the project also developed a Web control interface that allows the user to control the walking path of the robot through the Web, as shown in Figure 8:

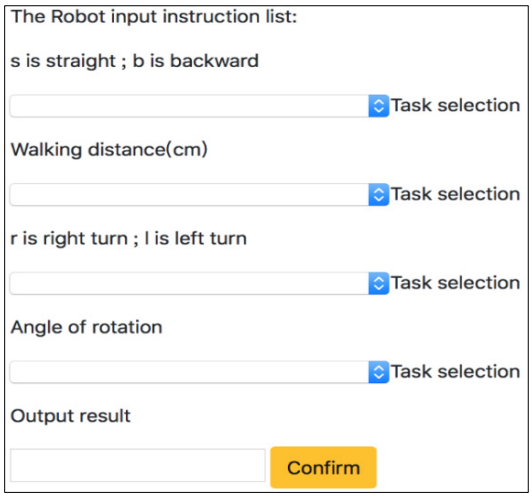


Figure 8: Manual specification of robot path

When the user issues an instruction, the system will put each value into the database and then convert it to json format output when needed, as shown in Figure 9

id	script
91	t200 r180
92	s50 l90
93	t100 r10
94	s50 l90
95	t10 r10
96	s50 l90
97	s50 l90
98	t100 r180
99	s10 l10
100	t50 r180
101	s200 r180
102	s100 l10
103	s50 r90

Figure 9: Store forward distance and turning angle

As shown in Figure 10, when robots execute Script, it would obtain forwarding distance and rotation angles from the database to execute tasks. When all assigned tasks are done, it would show the spent time for finishing all movements to notify users of time spent to complete cleaning so that users can more easily use it.

```
Today: Fri Feb  2 16:53:24 2018 16:53:24.457
Mission: t10 r10
Default speed 200mm/s

 | --- Start ---
 | - Straight walk -10 cm
 | - 0.494894 sec
 | - Turn 10 degrees
 | - 0.183419 sec
 | --- End ---

 | - It cost 0.747108 sec
Today: Fri Feb  2 16:53:25 2018 16:53:25.204
```

Figure 10: Executed Motions and Spent Time of Robots

3.3. *The Design of iBeacon Information Assist System*

Owing to smart robots can provide more human-oriented location-based service after combining with perception positioning system and have better market competitiveness in the future, this study new robot communication application service in view of the development of smart home and IoT. This study implements diverse simulation test by iRobot Create2 which is designed to control robots, applying iBeacon micro-location technology, compound multi-axis sensor technology and so on to develop smart robots for indoor positioning and path algorithm; for robots, the precision of positioning signals is a very crucial parameter.

3.3.1. *Experimental field planning and design*

To enable robots to obtain indoor coordinates and execute dynamic searching, this study constructs an IoT micro-location

system. Through the assistance of the iBeacon system, robots are able to inquire past moving track before executing tasks and transform the moving process to coordinate record to build up the spatial perception ability of robots; when robots calibrate signals and connect to system database. The robot can report its location to the system while forwarding and meanwhile report its location to the IoT micro-location system so that it can plan, correct and predict its paths.

Fig. 11 shows the experimental field. The experimental field is an 85m x 20m 2-dimensional space. A total of 14 Beacons are arranged to allow the robot to perform tasks that can be used for positioning applications.

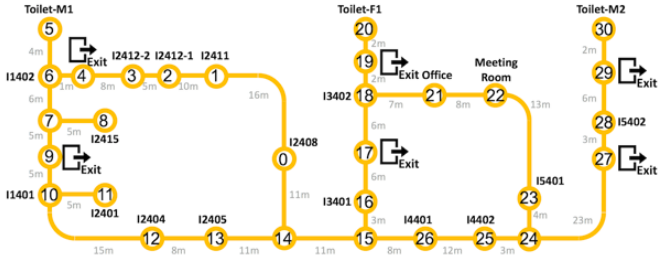


Figure 11: Experimental space for robotic indoor positioning applications

3.3.2. *Researches on Beacon Positioning*

The iBeacon messages detected by the robot contained various items of location-related information, including the UUID, the Tx Power, and so on (see Fig. 12)

```
pi@raspberrypi:~/iBeacon-Scanner- $ sudo python testblescan.py
ble thread started
----------
28:1d:2a:c4:ba:ac,5e5f9c2c626deea1145a5321c057319e,23150,1036,89,-71
4e:ab:8b:77:cd:49,0613ff4c000c0e004d219b3bf2d2bab0,21997,47890,-77,-49
f4:5c:89:c4:27:98,0100009827c4895cf40b02010607ff4c,16,523,0,-64
c8:69:cd:6c:34:64,01000064346ccd69c80b02011a07ff4c,16,513,0,-49
17:42:ec:af:bc:f8,bcafec42170f02011b0bff4c00090603,632,29288,32,-49
28:1d:2a:c4:ba:ac,5e5f9c2c626deea1145a5321c057319e,23150,1036,89,-62
4e:ab:8b:77:cd:49,0613ff4c000c0e004d219b3bf2d2bab0,21997,47890,-77,-51
f4:5c:89:c4:27:98,0100009827c4895cf40b02010607ff4c,16,523,0,-49
17:42:ec:af:bc:f8,bcafec42170f02011b0bff4c00090603,632,29288,32,-63
----------
```

Figure 12: Contents of iBeacon message

After obtaining RSSI and TxPower value between robots and Beacon, we can then use relative relation between RSSI and TxPower as the reference value of distance between the robot and Beacon; when ratio between RSSI and TxPower is less than 1.0, we use Eq. (3) to obtain distance, or we will use Eq. (4):

$$distance = \begin{cases} \left(\frac{RSSI}{TxPower}\right)^{10} & ,if\ \frac{RSSI}{TxPower} < 1.0 \quad (3) \\ \left(\frac{RSSI}{TxPower}\right)^{7.7095} \times 0.89976 + 0.111 & ,else\ (4) \end{cases}$$

Scan the iBeacon through the Raspberry Pi and calculate the distance, as shown in Figure 13, but since this method is easy to generate large fluctuations through real-time operation, it is necessary to increase the stability of the distance by accumulating balance, and plan two points iBeacon to be placed at 0.5m and 2.5m, 30 points are measured for each point, and the first 10 data are averaged, and then the last 20 data are corrected by the EWMA formula to obtain the distance. Figure 14 is the experimental results.

```
BLE thread started
MAJOR : 32 , MINOR : 14
TxPower : -59 , RSSI : -49
Ratio : 0.830508
Distance : 0.156114
--------*--------
pi@raspberrypi:~/iBeacon-Scanner- $
```

Figure 13: Raspberry Pi Scan iBeacon and Calculate Distance

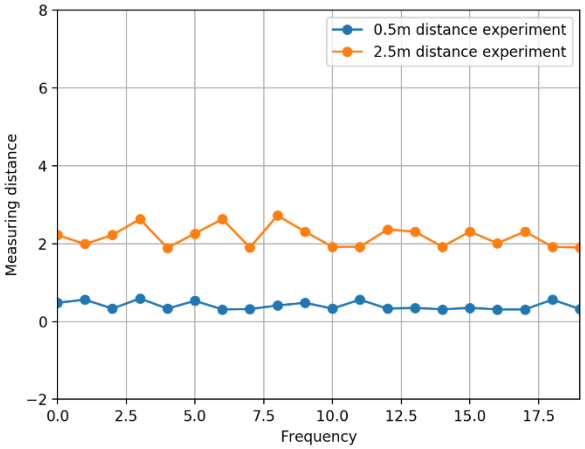


Figure 14: Revise the distance of iBeacon

3.4. *Robot positioning and map construction*

Safe walking is a challenge when robots enter an unknown environment. SLAM technology can be used to map assisted positioning by reading the value of the surrounding environment by the sensor.

3.4.1. *Robot Operating System*

With the warming of domestic robots in the past two years, autonomous positioning has become the primary threshold for robotic intelligence. To realize this function, many manufacturers have chosen to use ROS as the operating system of robots.

ROS is more like an application role as a communication intermediary software in the entire robot design architecture. At the same time, a complete system integration framework for robot application design based on an existing operating system can be developed in various programming languages such as C++ and Python. Through the functions provided by ROS, the use of SLAM can be made more convenient. Figure 15 shows the use of ROS to generate a simple picture through the virtual machine. It is necessary to constantly correct the map through the SLAM algorithm to make the map more accurate.

3.4.2. *SLAM map construction*

As shown in Fig. 16, SLAM comprises three components, namely localization, mapping and navigation. With advances in computer technology and sensing capabilities, SLAM has evolved into the method of choice for robots to detect unfamiliar environments and find their way through them. As shown in Fig. 17, various SLAM technologies are available, including VSLAM,

Wifi-SLAM and Rplidar SLAM. Among these technologies, Rplidar SLAM is one of the most commonly used and is hence also adopted in the present study.

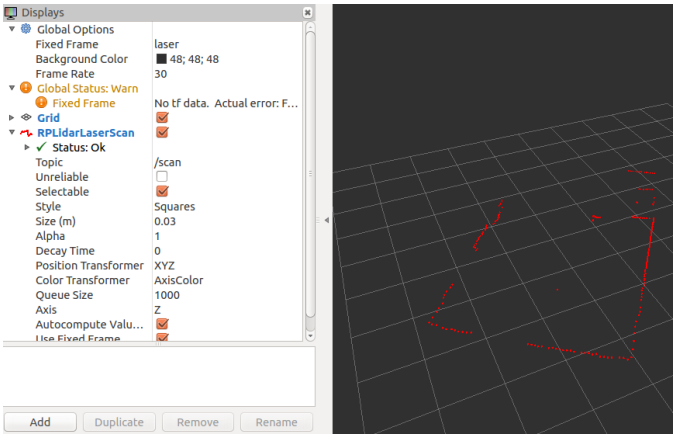


Figure 15: A Depicting map outlines through ROS

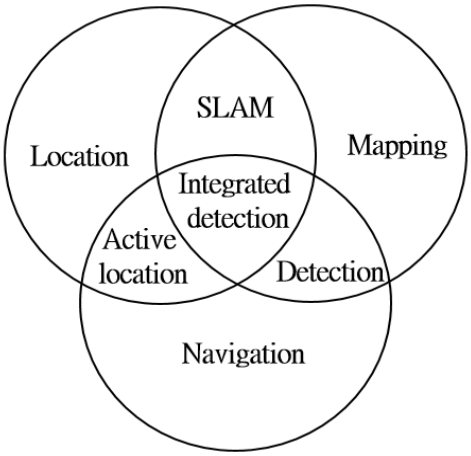


Figure 16: SLAM application diagram

VSLAM WiFi-SLAM Rplidar SLAM

Figure 17: Common SLAM technology

Rplidar Collect data repeatedly, filter noise and produce visual map, Rplidar A2 Typical rotational frequency is 10Hz (600RPM), can realize 0.9° angle resolution at typical rotational frequency, can adjust the rotation frequency in the range of 5Hz-15Hz with the user's demand.

3.4.3. *RPLIDAR Scan Value*

As shown in Figure 18, RPLIDAR is performed by Python to perform scanning, and the values of four angles are obtained, which are 0 degrees, 90 degrees, 180 degrees, and 270 degrees, and the results are shown in Table 6, can be applied to robot detection and The distance between obstacles.

Figure 18: Use Rplidar to get the value of the angle

Table 6: 4 angle values

Angle	Distance (cm)
0	17.6
90	103.3
180	13.4
270	20.0

3.4.4. *SLAM Algorithm*

In the SLAM algorithm, the Hector SLAM algorithm is used. The algorithm developed by Hector Labs in Germany was ported from the original RoboFram algorithm to the ROS package at the end of 2010. Using the 2D mesh method, the value of the laser scan is converted in the xy coordinates by the tf library, and possible position estimation and rendering obstacles. Repeated scanning uses a bilinear filtering method to continuously correct the map with low power consumption, immediacy and goodness. Figure 19 shows the working principle of the Hector SLM, and Figure 20 shows the experimental results of the Hector SLAM scanning virtual machine.

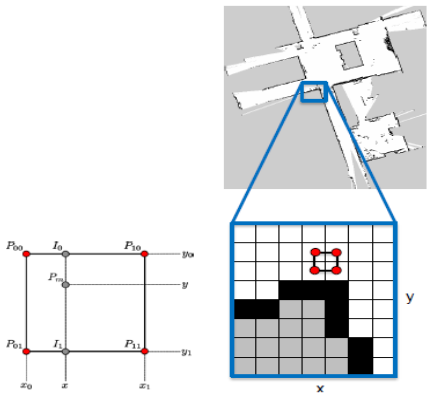


Figure 19: Hector SLAM works

3.5. *Robot Obstacle Avoidance*

To carry out research on robot obstacle avoidance, the robot is equipped with three ultrasonic sensors. As shown in Figure 21, the ultrasonic sensor can sense a distance of 2cm to 400cm, an accuracy of 0.3cm, and an angle range of 15 degrees. The robot can read the value of the sensor to determine whether it encounters an obstacle, such as Figure 22 shows.

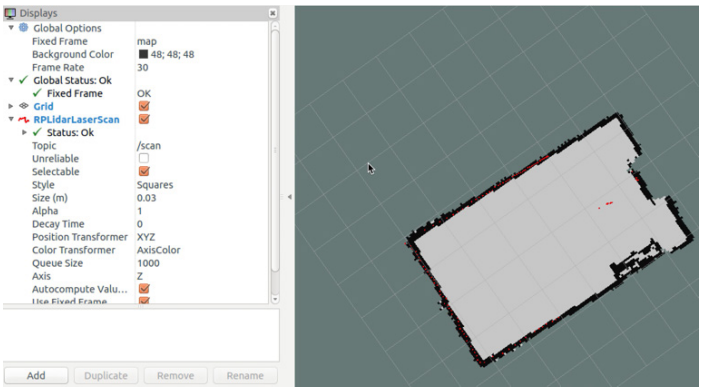


Figure 20: Logo

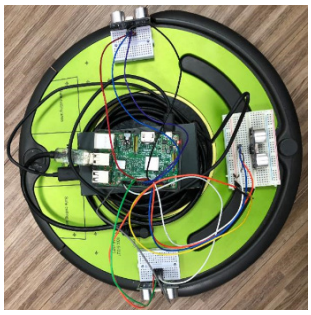


Figure 21: Install an ultrasonic sensor on the robot

```
pi@raspberrypi:~/Create2Control $ sudo python sensor_test.py
SensorNum:1 init
SensorNum:2 init
SensorNum:3 init
hinder:{'Front': 118.1, 'Right': 11.1, 'Left': 23.2}
hinder:{'Front': 118.1, 'Right': 11.1, 'Left': 23.2}
hinder:{'Front': 118.1, 'Right': 11.1, 'Left': 23.2}
```

Figure 22: Confirm that the ultrasonic sensor is working

The robot's ultrasonic sensor determines if an obstacle is encountered in front, and the robot will determine which side is longer by the left and right sensors to determine how to avoid the obstacle. As shown in Fig. 23, when the distance is less than 20 cm, the obstacle avoidance behavior is performed. When the sensor's value returns to normal, the robot will travel midway through the body to ensure that the robot completely bypasses the obstacle and rotates back to the original path.

```
hinder:{'Front': 22.4, 'Right': 115.1, 'Left': 81.9}
Car State: Stop
Car State: Turn Right
```

Figure 23: Turning with obstacles less than 20cm

To understand the execution situation of the obstacle avoidance action of the robot, Figure 24 and Figure 25 simulate and observe the execution process of the obstacle encountered by the robot, where the x-axis is time and the y-axis is the path of the robot to avoid obstacles. When the robot turns right the y axis is decremented by 1 , the y axis is incremented by 1 when it is turned left. Figure 24 simulates the robot encountering an obstacle at 2 seconds, then turning right around the obstacle and leaving the obstacle at 4 seconds. Figure 25 simulates the robot encountering an obstacle at 3 seconds, then turns left around the obstacle and leaves the obstacle at 5 seconds.

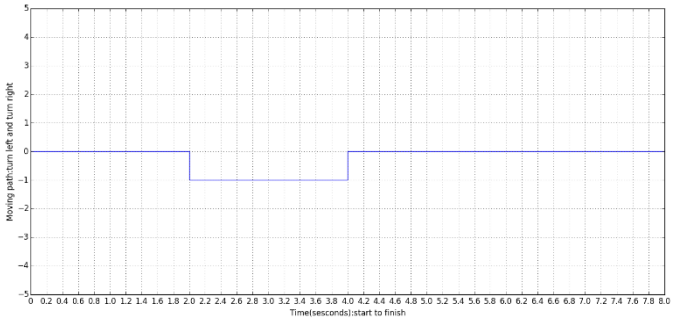


Figure 24: Simulate turn right bypass the obstacle

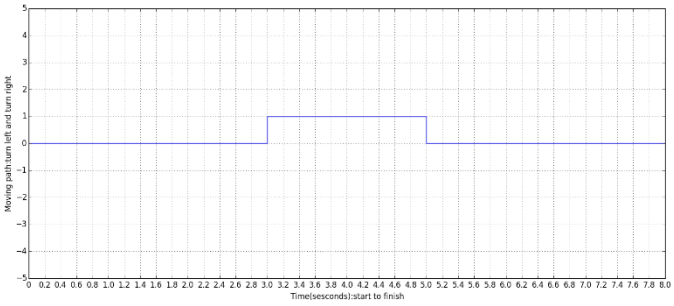


Figure 25: Simulate turn left bypass the obstacle

Let the robot perform the task of bypassing obstacles. From the experimental results, the robot encountered an obstacle within 2.4 seconds, began to perform the action of avoiding the obstacle, and left the obstacle within 3.8 seconds. Figure 26 is the result of the actual walking of the robot. The obtained time is brought into the program, and the result chart of the simulation is shown in Fig. 27, and the actual obstacle avoidance result chart is consistent.

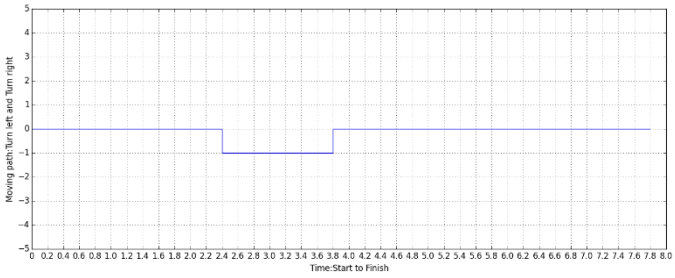


Figure 26: The actual obstacle avoidance result of the robot

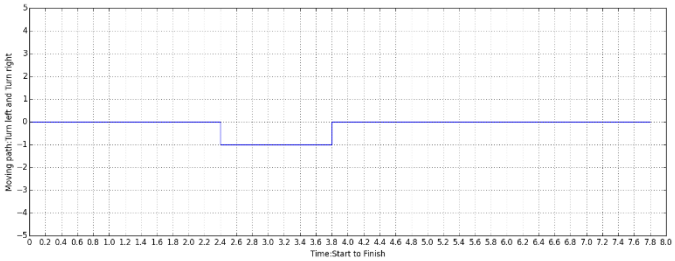


Figure 27: Simulate the results of actual obstacle avoidance

4. Conclusion

This study has developed an integrated development platform for indoor mobile robots consisting of a programmable iRobot Create 2 sweeping robot; a Raspberry Pi computer; a sensing system consisting of laser, magneto, accelerometer, gyroscope and

ultrasonic devices; and iBeacon technology. The path planning and positioning performance of the proposed platform has been evaluated experimentally and calibration equations have been proposed for both the linear and rotational motions of the robot. The use of SLAM technology to construct a map of the surrounding environment has been explained. Finally, the ability of the sweeping robot to perform automatic obstacle avoidance maneuvers has been demonstrated numerically and experimentally. In general, the results confirm the basic feasibility of the proposed framework and provide a useful basis for the future development of mobile robots for indoor smart network applications.

References

[1] Wengpeng Wang, Xin Wang "Research on robot indoor localization method based on wireless sensor network", 2015 International Conference on Advances in Mechanical Engineering and Industrial Informatics (AMEII 2015), May 2015.

[2] Md. Masum Billah, Md. Raisuddin Khan "Smart Inertial Sensor-based Navigation System for Flexible Snake Robot", IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), pp.1-5, Nov. 2014.

[3] Nieves Pavón-Pulido, Juan Antonio López-Riquelme, Juan Jesús Pinuaga-Cascales, Joaquín Ferruz-Melero, Raul Morais dos Santos "Cybi: A Smart Companion Robot for Elderly People: Improving Teleoperation and Telepresence Skills by Combining Cloud Computing Technologies and Fuzzy Logic ", EEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), pp.198-203, April 2015.

[4] Abhishek Pandey, Anirudh Kaushik, Amit Kumar Jha,Girish Kapse,"A Technological Survey on Autonomous Home Cleaning Robots", International Journal of Scientific and Research Publications, Vol. 4, No. 4, April 2014

[5] Ercan U. Acar, Howie Choset, Yangang Zhang and Mark Schervish, "Path Planning for Robotic Demining: Robust Sensor-Based Coverage of Unstructured Environments and Probabilistic Methods", The International Journal of Robotics Research, July 2003.

[6] Feng-Min Chang and Feng-Li Lian, "Inverse observation model and multiple hypothesis tracking for indoor mobile robots", 2014 IEEE International Conference on Automation Science and Engineering (CASE), pp. 1200-1205, 18-22 Aug. 2014.

[7] Zeyn Saigol, Bram Ridder, Minlue Wang, Richard Dearden, Maria Fox, Nick Hawes, David M Lane, and Derek Long, "Efficient Search for Known Objects in Unknown Environments Using Autonomous Indoor Robots", In Workshop on Task Planning for Intelligent Robots in Service and Manufacturing, IROS 2015. October 2015.

[8] Feng-Min Chang and Feng-Li Lian, "Inverse observation model and multiple hypothesis tracking for indoor mobile robots", 2014 IEEE International Conference on Automation Science and Engineering (CASE), pp. 1200-1205, 18-22 Aug. 2014.

[9] Zeyn Saigol, Bram Ridder, Minlue Wang, Richard Dearden, Maria Fox, Nick Hawes, David M Lane, and Derek Long, "Efficient Search for Known Objects in Unknown Environments Using Autonomous Indoor Robots", In Workshop on Task Planning for Intelligent Robots in Service and Manufacturing, IROS 2015. October 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 200-216 (2019)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Development of Application Specific Electronic Nose for Monitoring the Atmospheric Hazards in Confined Space

Muhammad Aizat Bin Abu Bakar[*,1], Abu Hassan Bin Abdullah[2], Fathinul Syahir Bin Ahmad Sa'ad[2]

[1]*Faculty of Engineering Technology, Universiti Malaysia Perlis, Kampus UniCITI Alam, 02100, Padang Besar, Perlis, Malaysia*

[2]*School of Mechatronic Engineering, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600, Arau, Perlis, Malaysia*

ARTICLE INFO

Article history:
Received: 20 August, 2018
Accepted: 04 February, 2019
Online : 15 February, 2019

Keywords:
Confined Space
Atmospheric Hazards
Electronic Nose
Purging
Pre-processing
Baseline Manipulation
Normality Test
PCA
SVM
RBFNN

ABSTRACT

The presence of atmospheric hazards in confined space can contribute towards atmospheric hazards accidents that threaten the worker safety and industry progress. To avoid this, the environment needs to be observed. The air sample can be monitored using the integration of electronic nose (e-nose) and mobile robot. Current technology to monitor the atmospheric hazards is applied before entering confined spaces called pre-entry by using a gas detector. This work aims to develop an instrument to assist workers during pre-entry for atmosphere testing. The developed instrument using specific sensor arrays which were identified based on main hazardous gasses effective value. The instrument utilizes multivariate statistical analysis that is Principal Component Analysis (PCA) for discriminate the different concentrations of gases. The Support Vector Machine (SVM) and Artificial Neural Network (ANN) that is Radial Basis Function Neural Network (RBFNN) are used to classify the acquired data from the air sample. This will increase the instrument capability while the portability will minimize the size and operational complexity as well as increase user friendliness. The instrument was successfully developed, tested and calibrated using fixed concentrations of gases samples. The results proved that the developed instrument is able to discriminate an air sample using PCA with total variation for 99.42%, while the classifier success rate for SVM and RBFNN indicates at 99.28% for train performance and 98.33% for test performance. This will contribute significantly to acquiring a new and alternative method of using the instrument for monitoring the atmospheric hazards in confined space.

1. Introduction

This paper is an extension of work originally presented in IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA 2017) in title of Electronic Nose Purging Technique for Confined Space Application [1]. A confined space is large enough for workers to enter and perform work. It has a limited means of entry or exit and is not designed for continuous occupancy because it's could contribute towards atmospheric hazard accidents. Accidents do happen sometime in these areas and usually involve human fatalities or death. The Occupational Safety and Health Administration (OSHA) and National Institute of Occupational Safety and Health (NIOSH) state that the presence of atmospheric hazards in confined space are serious environmental problem that threatens the industry operation and safety of the workers [2].

The hazards in confined spaces can be classified into two categories which are physical hazards and atmospheric hazards [3]. The physical hazards can be visualized and avoided by taking initial safety precautions. The example of physical hazards includes unstable materials, moving parts of machinery, falling objects, a slippery surface and noise. Atmospheric hazards are more dangerous compared to physical hazards as they are unseen and come from oxygen deficiencies, hazardous gases, dust and welding fumes. The hazardous gases can interfere with the human body's ability to transport and utilize oxygen as well as cause negative toxicological effects. Usually the atmospheric testing in confined space is carried out during pre-entry by authorised person using gas detectors. The atmospheric hazards are rated into three stages which are High, Moderate and Low.

The atmospheric hazards workers are exposed to in confined spaces normally involve oxygen (O_2) too low or high or the

[*] Muhammad Aizat Bin Abu Bakar, Email: ijatbakar@gmail.com

presence of flammable and toxic gases [4]. The main flammable gas is methane (CH_4) while the toxic gases include hydrogen sulphide (H_2S) and carbon monoxide (CO). Before workers enter the confined space, a pre-entry for atmospheric testing is conducted by the authorised person for safety requirement. The atmospheric hazard conditions must be monitored before and while workers perform their activities inside the area. The hazards can cause serious health problems or death to the workers if not monitored properly [5].

At present, the pre-entry for atmospheric testing is done by using a direct-reading gas detector that need to carry by the tester towards specific location [3]. The worker (tester) is exposed with the atmospheric hazards directly during pre-entry testing activity is carried out. Even do, there are multi gas detector in the market, but the acquired measured data still not reliable due to the purging system that cause low repeatability [6]. The detector also shows the real time measurement at specific location only which not represent the whole confined space environment. Therefore, there is a need for a system like e-nose that is able to measure the hazardous gases and predicts the atmospheric hazards in confined space with high accuracy and repeatability [7].

An electronic nose (e-nose) is an instrument which comprises an array of electronic chemical sensors with partial sensitivity, an appropriate pattern recognition system and capable of recognising simple or complex odours [8]. Development of this instrument over the past decades is significant for its possible applications and achievements [9]. The application includes food quality assurance, work safety, medical diagnosis, plant disease detection and environmental monitoring [10]. An e-nose has shown a good potential for detecting and monitoring atmospheric hazards present in the confined space which could contribute to deadly accidents. A good e-nose must be able to produce the same pattern for a sample on the same array to maintain its repeatability [7].

The main aim of this work is to develop an instrument to assist workers during pre-entry for atmosphere testing in confined space (i.e. hospital mechanical room) and which address the following:

i. To investigate and identify the atmospheric hazards main hazardous gases.

ii. To design and fabricate e-nose system with multimodal sensor detection.

iii. To integrate the system with optimum self-purging.

iv. To test and validate the functionality of fabricate system in laboratory and field environment.

2. Atmospheric Hazards

The atmospheric hazards can only be detected by sense of smell. The main atmospheric hazards in confined space are oxygen (too much or too little), flammable and toxic atmosphere [4].

2.1. Oxygen Atmosphere

The normal air in the atmosphere is approximately composed of 21% oxygen and 79% nitrogen [4]. The oxygen deficiency in confined space may happen when metals rust, combustion engines run for a period time is replaced by other gases (i.e. welding gases) or used by micro-organisms (i.e. fermentation vessels). The enriched or over limit of oxygen in confined space is also dangerous in that it increases the risk of fire or explosion. Materials would quickly and easily burn when there is a high level of oxygen. Table 1 shows the potential effects of oxygen deficient and enriched atmospheres [11].

Table 1: Potential effect of oxygen enriched and deficient in atmosphere

Oxygen content	Effects and symptoms
> 23.5%	Oxygen enriched, extreme fire hazard
20.9%	Oxygen concentration in normal air
19.5%	Minimum permissible oxygen level
15% to 19%	Decreased ability to work strenuously may impair coordination and may cause early symptoms for persons of coronary, pulmonary or circulatory problems
10% to 14%	Respiration further increases in rate and depth, poor judgment, blue lips
8% to 10%	Mental failure, fainting, unconsciousness, ashen face, nausea and vomiting.
6% to 8%	Recovery still possible after four to five minutes. 50% fatal after six minutes. Fatal after eight minutes
4% to 6%	Coma in 40 seconds, convulsions, respiration ceases, death

2.2. Flammable Atmosphere

Three elements that termed the fire triangle are necessary for a fire or explosion to occur in confined spaces which are oxygen, flammable material (gas or fuel) and source of ignition (spark or flame).

For flammable material, Figure 1 illustrates the relationship between Lower Explosive Limit (L.E.L) and Upper Explosive Limit (U.E.L). The ignition sources of fire may start from open flames, welding arcs, lightning, sparks from metal impact, arcing of electrical motor or a chemical reaction [4]. Several industrial processes that generate static charge such as steam cleaning, purging and ventilation also have potential of fire hazards.

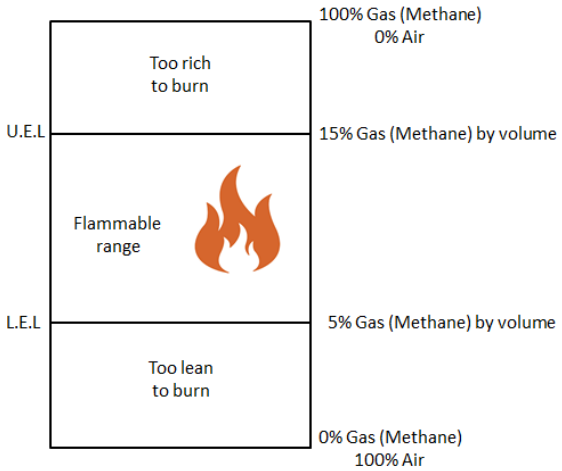


Figure 1: Lower Explosive Limit (L.E.L) and Upper Explosive Limit (U.E.L)

The L.E.L is the critical point at which ignition or explosion occurs [11]. Once the L.E.L is reached, the danger of a fire or

explosion is continuous all the way to U.E.L. For example the L.E.L of methane (CH_4) is 5% by volume and the U.E.L is 15% by volume. When a confined space reaches 2.5% of methane by volume this would be equal to 50% L.E.L which means that 5% methane by volume would be 100% L.E.L. Between 5 to 15% by volume, a spark could cause an explosion. Different gases have different percentage by volume concentrations to reach 100% L.E.L.

2.3. Toxic Atmosphere

The main toxic gases in confined space are carbon monoxide (CO) and hydrogen sulphide (H_2S) [4]. It may leak from products that are stored due to improper procedure or poor ventilation. The activities inside the area such as welding, painting, scraping and sanding may produce the toxic gases. The toxic fumes produced from nearby activities may flow and accumulate inside the area.

Workers expose to toxic gases in the atmosphere may be injured or killed by the gas. At certain concentrations, some toxic gases become too dangerous. At such levels, even a brief exposure can cause permanent health effects such as brain, heart and lung damage or the substance may render workers unconscious so that they cannot escape from the confined space. The Life-Threating Effects of carbon monoxide and hydrogen sulphide are listed in Table 2 and Table 3 respectively [11].

Table 2: Effect of carbon monoxide exposed period

Exposure (ppm)	Time	Effects and symptoms
35	8 hour	Permissible Exposure Level
200	3 hour	Slight headache, discomfort
400	2 hour	Headache, discomfort
600	1 hour	Headache, discomfort
1000 to 2000	2 hour	Confusion, discomfort
2000 to 2500	30 minutes	Unconsciousness
4000	> 1 hour	Fatal

Table 3: Effect of hydrogen sulphide exposed period

Exposure (ppm)	Time	Effects and symptoms
10	8 hour	Permissible exposure level
50 to 100	1 hour	Mild eye and respiratory irritation
200 to 300	1 hour	Marked eye and respiratory irritation
500 to 700	½ to 1 hour	Unconsciousness, fatal
> 1000	1 minutes	Unconsciousness, fatal

3. Fundamental of Electronic Nose System

3.1. Electronic Nose

The e-nose as a device to mimic the discrimination of the mammalian olfactory system for smell is introduced in 1982 [12]. Initially the instrument used three different Metal Oxide Semiconductor (MOS) gas sensors to identify several chemical volatile compounds by using the response steady state signals. Then Gardner [13], described the e-nose development comprise of: (1) a matrix sensor to simulate the receptors of the human olfactory bulb, (2) a data processing unit that performs the same function as the olfactory bulb and (3) a pattern recognition system that would recognize the olfactory patterns of the substance being tested, which is a function performed by the brain in the human olfactory system. Figure 2 shows the similarity between human olfactory system and the e-nose system.

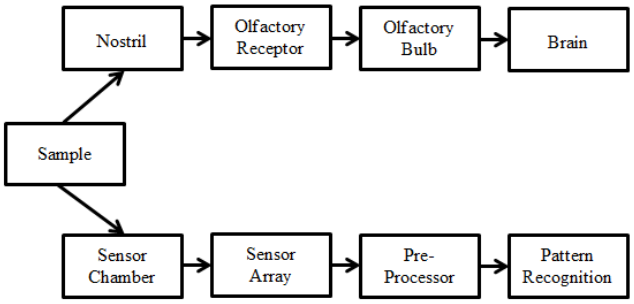


Figure 2: Similarity between human olfactory system and e-nose system

3.2. *Sample Handling*

The sample handling is a process to deliver the air sample to the sensor array. In the detection process, a sensor chamber is used to locate the sensor array where it interacts with the air sample. The data processing unit and pattern recognition algorithm will process the acquired data and classify them accordingly. The standard e-nose operating procedure includes determination of sample preparation and sampling method. The sample preparation procedure is determined by the type of sampling either static or dynamic method [14].

For static sampling, the sensor array and the sample is located at the same place in the chamber. The reading of sensor array is taken from the sample after the headspace in the chamber is homogenised. For dynamic sampling, the sensor array and the sample are at different or separate locations. The chamber is designed to locate a set of sensor arrays in one cavity and the sample is delivered into another cavity in order to be exposed to the sensors. This is also known as the Odour Capturing Module (OCM). The static sampling method is suitable to implement at a laboratory and not practical for outdoor use, while OCM is suitable to implement in both environments. However, the OCM must be purged to clean the chamber cavity after every sampling process to ensure sensor stability and repeatability for the next sampling process to produce consistent output readings [15].

Sample handling is a critical step affecting analysis by e-noses. The quality of the analysis can be greatly improved by adopting an appropriate sampling technique. To introduce the volatile compounds present in the headspace (HS) of the sample into the e-noses detection system, several sampling techniques have been used [16]. Various types of systems have been developed and are used to gather an air sample for analysis such an electric air pump which is used to suck the outside air sample into the sensor chamber [17]. A tube connected to a dedicated pump is generally used to direct the air sample into the e-nose. Other than that, the air sample delivery system was designed to provide seamless control over the operation by using a fan [18].

3.3. Sensor Chamber

A sensor chamber is used to accommodate the sensor array's interaction with the air sample. The chamber must be properly developed for optimum sensor response measurement. The design will emphasize optimum sensor response, stability, reproducibility and repeatability [19]. The air sample flow inside the chamber should be homogenous with low velocity to minimise the recirculating zones and stagnant volume [7]. This homogenous

flow is to ensure that the sensor arrays are exposed simultaneously to the air sample for optimum response measurement. The sensors positions in the chamber can be either in a series or parallel to the air sample. The different sensors position from the chamber inlet will cause a time shift in each sensor response to the air sample. Since the distance between the sensors is small, so the time shift effect can be ignored [20].

The symmetric structure will enable sensor arrays to respond simultaneously to the incoming air sample. However, this structure size is quite large to accommodate the sensor arrays in one area. The chamber development must focus on the structure geometry, material selection, type of flow and sensors position which will optimise its performance. The geometry of the chamber should be symmetrical to enable the air sample to be exposed to the sensor arrays efficiently. The flow of air sample is usually characterized using dimensionless ratio known as Reynolds number (*Re*). The *Re* is a measure of the kinetic forces to the viscous forces in a flowing fluid. The Navier-Stokes formula shown in Equation 1 is used to solve for air sample flow characteristics [21].

$$\mathrm{Re} = \rho du / \mu \tag{1}$$

Where,

- ρ = density of the fluid (kg/m^3)
- d = hydraulic diameter (m)
- u = fluid mean velocity (m/s)
- μ = dynamic viscosity of the fluid (kg/ms)

3.4. Gas Sensor

The gas sensor is used by the e-nose to interact with the air sample inside the sensor chamber. The interaction will cause a change in certain chemical and physical properties known as sensor responses that will convert into electrical signal. These sensor responses will be acquired by a laptop computer that links wirelessly with the e-nose to be used by an off-line signal processing method. The instrument sensors sensitivity should be suitable with the application's volatile compounds [12]. The sensors are also selected based on their characteristics which include fast response, stability, reproducibility and reversibility.

The principles of gas sensors are optical, thermal, electrochemical and gravimetric [22]. Most e-nose devices use electrochemical sensors to detect chemical substances to response. The operating principles are based on changes in the conductivity of sensing material by either adsorption or absorption of the gaseous molecules. The sensor conductivity depends on the material used and is relative to the sample concentration. The gas sensors performance is measured by several criteria and their behaviour include sensitivity, selectivity, stability, detection limit, responses time, recovery time, life cycle and operating temperature [23]. All these criteria are very important to take into consideration during the selection of the gas sensors. The MOS gas sensor was widely use selected based on its stable response in performance.

3.5. Microcontroller System

The e-nose normally uses an embedded controller which is a microcontroller with embedded software for operation, data acquisition and classification [24]. The control and system software is embedded in the instrument memory that may perform concurrently during operation. A microcontroller is a component that is used to control the sampling process for acquiring the sensor response signals to be processed by a personal or laptop computer [25]. The microcontroller has good processing capabilities and a flexible interface to the instrument's components which make it the ideal choice as the controller. The microcontroller memories are RAM, Flash ROM and EEPROM. The Flash ROM memory is where the program is stored, also called program memory. RAM is used for the temporary data during run-time. The EEPROM is used for data memory that needs to be retained during power failure. This microcontroller memories ability makes it suitable for e-nose operation that consist multimodal sensor to send the signals concurrently.

Currently, the dsPIC33 microcontroller type from Microchip Technology Inc. as the embedded controller was used in e-nose system development [26]. The microcontroller was designed using Surface Mounted Technology (SMT) component to improve the performance by reducing the signal to noise ratio (SNR). The microcontroller also converts the acquired analogue signal to digital by using an on-board Analogue-to-Digital Converter (ADC).

3.6. Signal Conditioning

The e-nose signal conditioning frame work as shown in Figure 3 consists of an interface circuit, conditioning circuit and filter [27]. The interface circuit measures the sensor responses and converts it into output voltage that varies with conductive change in the sensing element [28]. A conditioning circuit is used to buffer and amplify the signal to suit the microcontroller requirement. A voltage follower is used as a buffer to isolate the signal and as impedance matching of the sensors output. The noise is removed from the signal by using a passive Low Pass Filter (LPF) because it is good for removing a small amount of high frequency noise (>2 kHz).

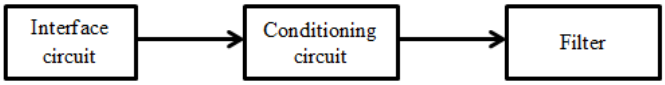


Figure 3: The signal conditioning frame work

3.7. Data Processing

The e-nose will produce a time-series sensor response corresponding to the air sample. The sensor response as data also depends on ambient air, temperature and humidity. The process consists of pre-processing, dimension reduction, pattern recognition and results as shown in Figure 4.

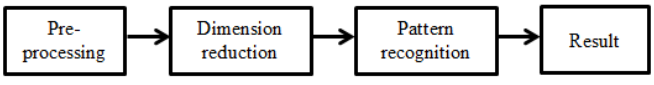


Figure 4: Electronic nose data processing block diagram

The chemo-metric processing technique that relates mathematical and statistical is used to extract the relevant information from the data [29]. This method uses multivariate analysis to discriminate the data simultaneously. The process uses either a statistical or biological classification method approach to classify or cluster the data qualitatively or quasi quantitatively. The

analysis methods are divided into parametric, non-parametric, supervised and unsupervised [30]. A parametric pattern is when the data are Gaussian and the distribution feature is normal. Supervised method develops a classification model using known data while unsupervised method uses unknown data for the development.

3.8. Pre-processing

The e-nose raw data is highly susceptible to noise, uncertain value, and inconsistency pattern. The quality of the data affects the classification model result. The data needs to be pre-processed to improve its quality which will in turn improve the classification process [31]. The process is critical for data processing that involves preparation and transformation of the initial raw data.

3.9. Feature Selection

The e-nose time series acquired data consists of dynamic or transient and steady state sensor response [32]. Figure 5 shows the transient and steady state sensor response. Feature selection is used to select a certain region of the sensor response that contains relevant sample information. Most of the e-noses extract the steady state region from the sensor response for the pattern recognition process.

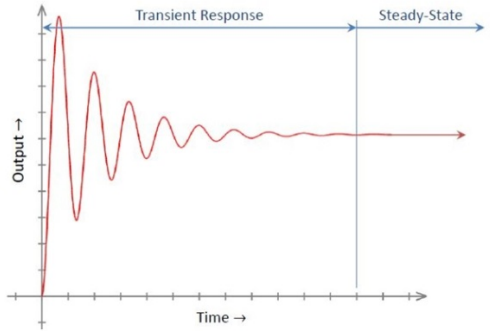


Figure 5: The transient and steady state sensor response

3.10. Baseline Manipulation

The baseline manipulation is a method that is based on the difference of the sensor response value between reference and sample [33]. The reference normally is the ambient air or nitrogen gas. Four baseline manipulation methods are commonly employed which are difference, relative, fractional and logarithm. The difference method is a widely used by directly subtracts the baseline and can be used to eliminate additive drift from the sensor response. Relative manipulation divides by the baseline, removing multiplicative drift and generating a dimensionless response. Fractional manipulation subtracts and divides by the baseline, generating dimensionless and normalized responses. The advantages of these processes will enhance the quality of data which contains the relevant sample information.

3.11. Normality Test

The procedure in assessing a sample of data has a normal distribution are graphical methods (histogram, boxplots and Q-plots), numerical methods (skewness and kurtosis indeces) and normality test [34]. A normality test is a statistical process used to determine if a sample or any group of data fits a normal distribution or not to determine suitable types of data analysis.

A normality test can be performed mathematically for example the Shapiro-Wilk (SW), Kolmogorov- Smirnov (KS) or Anderson-Darling (AD) test [34]. The power of test method proportionally increased with sample size and level but was still low for a small sample. The Kolmogorov- Smirnov test is the best with high sensitivity in rejecting null hypothesis [35]. The Shapiro-Wilk is suggested being to be used for a normality test, but Kolmogorov-Smirnov is popular [36].

Basically the data skewness and outlier's features indicate non-normality distribution. The data non-normality distribution needs to be transforming in order to process using the parametric classification method [35]. But the alteration on the original non-normality data will create curvilinear relationships that will complicate the interpretation. The data should us a non-parametric classification method when the distribution is not normal.

3.12. Multivariate Statistical Analysis

The Multivariate Statistical Analysis is used to discriminate and classifier sample. It build discriminate and classifier models. Some of the multivariate analysis techniques are Principal Component Analysis (PCA) for discriminate and Support Vector Machine (SVM) for classifier. This technique is described in the following sub section.

3.13. Principal Component Analysis

The PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. The PCA will transform the number of original data with correlated variables (number of sensors) into uncorrelated variables (Principal Components, PC) and loadings or weight of each original variable of variance percentage.

Traditionally, PCA is performed on a square symmetric matrix. It can be a pure sum of squares and cross product (SSCP) matrix, covariance matrix (scaled sums of squares and cross products) or correlation matrix (sums of squares and cross products from standardized data). The SSCP and covariance do not differ, since these objects only differ in a global scaling factor. A correlation matrix is used if the variances of individual variants differ much or if the units of measurement of the individual variants differ. Basically, PCA can be represented in five steps [37]:

i. Normalize the data by subtracting the mean
ii. Calculate the covariance matrix
iii. Calculate the eigenvectors and eigenvalues of the covariance matrix
iv. Choose components and form a feature vector
v. Derive the new data set

The PCA result is shown by a two-dimensional (2D) or three-dimensional (3D) graphical plot that contains the original data's important information and shows the PC percentage and group cluster which is being analysed visually. The PCA is used as the dimension reduction technique as it is the most widely used in e-nose data analysis.

3.14. Support Vector Machine

The SVM is a supervised learning model that is based on statistical learning theory for data classification and regression

[38]. The SVM training used data with known class when designing a linear classifier hyper plane model for separate group. SVM is targeting clear maximum margins between the hyper plane and closest training. In addition, the SVM optimized classification decision function is based on structural risk minimization in order to avoid over fitting.

In more complicated for non-linear cases, SVMs employ the kernel trick where a positive definite kernel function is used to map the input data into a high dimensional transformed feature space and this method regularly used by the e-nose community that used multisensory system [39]. The SVM mathematical modelling and algorithm can be referring in [40]. Although SVMs were initially developed for solving two-class problems, but it is also can be easily adapted to tackle multiclass problems. The two common techniques consist of:

i. One per class (OPC). This method is also known as "one against others". OPC is simple and results in reasonable performance. K SVM classifiers are trained, each of which separates one class from the other (K−1) classes. For each measurement, x, to be classified, K SVM decision outputs fk(x), $1 \leq k \leq K$ are obtained. The class of measurement x, j, is determined as fj(x)/fk(x) > 1 for $(1 \leq k \leq K) / = j$.

ii. Pairwise coupling (PWC). This method is also known as "one against one". This method trains (1/2)K(K−1) binary SVM classifiers. To classify a measurement, PWC combines the scores of these (1/2)K(K−1) classifiers. Each of the (1/2)K(K−1) binary SVM classifiers provides a partial decision for classifying a measurement. There are different methods of combining the obtained classifier and the most common is a simple voting scheme. When classifying a new instance each one of the base classifiers casts a vote for one of the two classes used in its training.

Since SVM does not require any estimation of statistical distributions of classes to accomplish the classification task, the application is widely used in many fields. For examples, the classifier performance is excellent in determination of green tea quality grades [41] and identification of selected features on Mexican coffee [42].

3.15. Artificial Neural Network

The non-linear e-nose sensor responses have to be converted to linear before they are classified using linear classifier methods. However, high volume of data from complex instrument's sensor responses is difficult to produce an efficient classification model. The Artificial Neural Network **(**ANN) has good learning, generalization and noise tolerance that is suitable for the e-nose non-linear data [43]. It has the ability to learn the complex interaction between the array of sensors and the air sample [44]. The network's ability to compensate the moderately sensitive sensors towards the air sample has improved the classification success rate.

The ANN is a mathematic model in which the nodes are interconnected with each other with weights and biases to form the network architecture. The architecture depends on the mathematical analysis of task criteria. The common network structure uses fully connected three-layer architecture models which are made of input, hidden and output nodes. Figure 6 shows that an ANN model consists of the input layer, hidden layer and output layer.

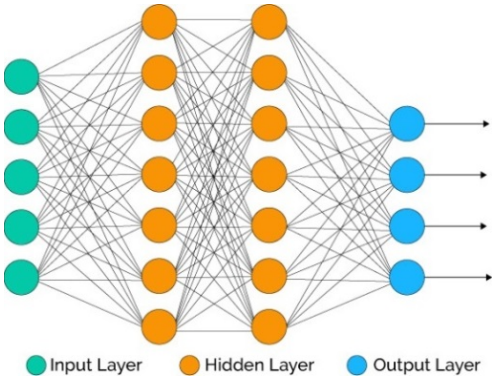


Figure 6: Basic neuron model showing the components and connections

Each neuron has a few weighted inputs and finally generates one output. The output is a function on these weighted inputs. Various types of functions can be used to calculate the output in a neuron. One type of neuron is called a perceptron which takes a vector of real-valued inputs and calculates a linear combination of these inputs. It outputs 1 if the result is greater than some threshold and -1 otherwise. Given inputs $x_1 ... x_n$, this perceptron function is shown in Equation 2 till Equation 5:

$$0(x_1 .....x_n) = \begin{cases} 1, & if\ w_1x_1 + w_2x_2 + ... + w_nx_n \geq w_0 \\ -1, & Otherwise \end{cases} \quad (2)$$

Where,

- w_0 : a threshold
- w_i : real-valued constant weights
- x_n, : inputs

An unthresholded perceptron is a linear unit, whose output is:

$$0(\bar{x}) = \bar{w}.\bar{x} \quad (3)$$

Where,

- $\bar{w}$: weights vector
- $\bar{x}$: inputs vector.

Another commonly used neuron is sigmoid unit, whose output is:

$$0 = g(\bar{w}.\bar{x}) \quad (4)$$

Where,

- g: sigmoid function

The logistic form of sigmoid function is:

$$g(y) = \frac{1}{1+e^{-y}} \quad (5)$$

A sigmoid unit is equal to a perceptron. Firstly, it calculates the linear combination of its inputs and then applies a threshold to the result. However, this threshold is continuous comparing with a

perceptron with a discontinuous threshold. The output of sigmoid function is between 0 and 1 (or between -1 and 1) and increases monotonically with its inputs. The sigmoid unit maps a very large input domain (-∞,∞) onto a small range of outputs (0,1) or (-1,1).

A particular transfer function is chosen to satisfy some specification of the problem that the model is attempting to solve [45]. During training, the model learns from the input neuron and gradually adjusts its weight to reflect the desire outputs [46]. The network model is tested by using the randomised unused testing data. The process uses the unknown data to test the network's generalisation ability of the classification model. The performance of the network model is given by the Equation 6.

$$Classification\ success\ rate = \frac{Correct\ classified\ data}{Total\ data} \times 100 \quad (6)$$

Types of ANNs refer to the model learning methods: supervised, unsupervised and reinforcement. These include Multilayer Feed-Forward Perceptron (MLP), Fuzzy ARTmaps, Kohonen's Self-Organizing Maps (SOMs), Probabilistic Neural Network (PNN) and Radial Basis Function Neural Network (RBFNN). The MLP, PNN and RBFNN are supervised types of classifiers that are normally used for e-nose applications [47]. The classification results using RBFNN is the most stable with a high classification success rate [48].

3.16. Radial Basis Function Neural Network

The RBFNN consists of three layers includes input layer, hidden layer and output layer. The method approximation capabilities are based on the superposition of local models of the response system. The output layer only computes a linear combination of the activation of the neurons in the hidden layer. The input output multivariate relationship is given by using Equation 7 and also can be written in matrix form in [49]:

$$y_{ij}(x) = \sum_{f=1}^{K} \sum_{i=1}^{N} w_{mj} G(\| x_i - c_m \|) + b + e_{ij} \quad (7)$$

Where,

- y_{ij} is *i,j-th* element of the output matrix $Y_{N,K}$
- N is the number of observations (data)
- K is the number of outputs in the RBFNN (or responses)
- If denote M as the number of hidden neurons (or the RBF canter's), w_{mj} is *m,j-th* element of the weight matrix $W_{M+1,K}$, m = 1, 2, …, M
- M+1 is the number of hidden neurons add bias (b)
- X_{NM} is the input patterns
- c_m is the square centroids matrix M x M
- E_{NK} is the matrix of residuals of e_{ij}
- $Y_{NK}(x)=d_{ij}$ is the process output

The activation of every neuron depends on the distance of the input vector to the prototype represented by the radial basis function and neuron parameter. The radial activation function network provides a nonlinear method of interpolating between numbers of different areas, time-series estimation and classification. The method always converges at the same point when trained with the orthogonal least squares algorithm. The advantage of the classifier is that the network has no local minima problem [50].

4. Electronic Nose Development

4.1. System Structure

The e-nose system structure includes the sensing module, signal conditioning, microcontroller and embedded software. The instrument peripherals (i.e. an air pump, keypad, graphical Liquid Crystal Display (LCD), purging system, power supply and wireless Radio Frequency (RF) communication are controlled by the microcontroller) as shown by Figure 7.

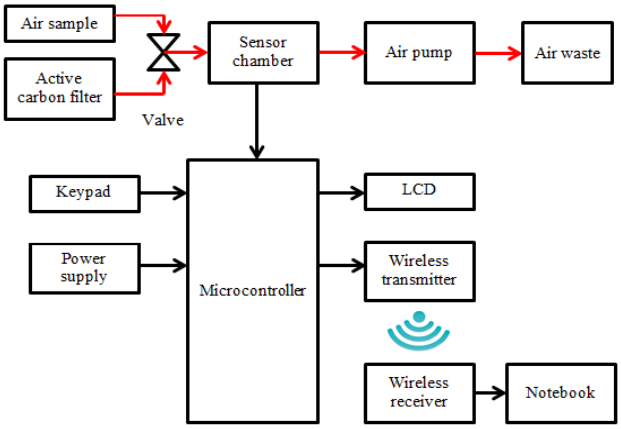


Figure 7: The e-nose block diagram

4.2. Sensing Module

The sensing module consists of sensor arrays for gas detection and a sensor chamber to locate the sensor arrays. Each gas sensor should have a different sensitivity profile over a range of hazardous gases for the confined space application. The process had selected suitable sensors with varying sensitivity to generate the responses known as smell prints. The process had reduced the instrument size, cost of the sensors, power consumption and will increase classification performance.

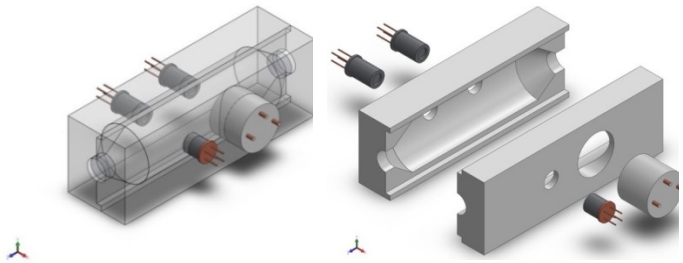
The gas sensors performance is measured by several criteria such as sensitivity, selectivity, stability, detection limit, response time, recovery time, life cycle and operating temperature [23]. All these criteria are very important to consider during selection of the gas sensors. The developed e-nose sensing module used four selected MOS gas sensors that are normally and widely used for environmental monitoring [51]. The sensors were selected based on each sensor varying sensitivity and selectivity that generated a smell print profile corresponding to the air sample. The selections were based on the confined space main hazardous gases effective value with different sensitivity manufactured by Figaro Inc. and Synkera Technologies Inc. as listed in Table 4. The sensors are able to detect main hazardous gases effective value to effect human for oxygen at between 0-30%, carbon monoxide at 35 ppm, hydrogen sulphide at 20 ppm and methane 5%. For monitoring the environmental conditions, the temperature and humidity sensor (SHT75) from Sensirion company was also located under the developed instrument due to its characteristics and ability (refer to appendix A). The sensor uses two wires through Serial Peripheral Interface (SPI) communication to communicate with the embedded controller.

Table 4: E-nose sensors selection

Category	Target Parameter	Sensor	Sensitivity (ppm)	Sensitivity (%)	Operating Temperature (°C)
Oxygen	Oxygen	Sk-25F	-	0 to 30	-10 to 50
Toxic	Carbon Monoxide	TGS 2442	30 to 1000	-	-10 to 50
	Hydrogen Sulphide	PN 714	1 to 100	-	-10 to 50
Flammable	Methane	TGS 2612	-	1 to 25	-10 to 40

The e-nose sensor chamber was designed and simulated using Computational Fluid Dynamic (CFD) SolidWorks version 2012 software. The SolidWorks software was used to create the Three Dimensional (3D) model of the chamber structure. Then the CFD was used to simulate and analyze for optimum air sample flow inside the chamber cavity to expose to the sensors. The analysis was based on the pressure, velocity and Reynolds (*Re*) number that were generated during the simulation.

The simulation finding was used to enhance the sample flow optimization, structure geometry, material and reduce the chamber dead zone [52]. The geometry should be symmetrical for efficient sample flow inside the chamber cavity. The simulation analyzed the best sensor position between series and parallel chamber design for optimum sample flow exposure to the sensor array and the series type was chosen shown in Figure 8.



(a) (b)

Figure 8: (a) Isometric and (b) explode view for chamber design

4.3. Signal Conditioning Board

The e-nose MOS gas sensors were attached to the signal conditioning board. The board is used to enhance the sensor responses signal suitable for the data acquisition by the microcontroller. The unit consists of the interface circuit, amplifier, filter and Analogue-to-Digital Converter (ADC) as shown in Figure 9.

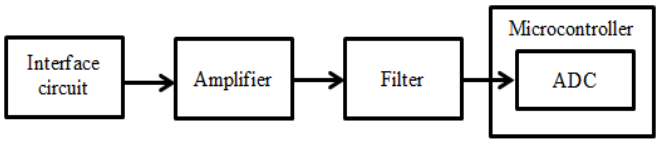


Figure 9: The signal conditioning block diagram

The interface circuit will measure the sensor responses and convert it into a constant current circuit. The sensor heater was used to increase the sensor temperature to around 300°C and the heater circuit is including in the sensor. The sensor response was calculated by using Equation 8.

$$R_s = \frac{V_C - V_{RL}}{V_{RL}} \times R_L \quad (8)$$

Where,

- R_s is the sensor resistance or responses
- V_c or V_{cc} is the circuit voltage
- V_{RL} is the voltage load resistor or voltage output (V_{out})
- R_L is the load resistor

The R_L value used to limit the sensor output voltage and V_{RL} up to five volts. The R_L were using 10 KΩ with 1% accuracy to enable the output voltage range suitable for the microcontroller. The signals from sensors are in analogue form and voltage divider rule are applied to the signal conditioning circuit. The sensor principle is impedance with low range (refer to appendix A) and the MCP602 Integrated Circuits (IC) unity gain voltage follower from Microchip Technology Inc. was used as the signal amplifier. The amplifiers high input and low output with good impedance matching will minimise the loading effects to the microcontroller ADC.

A passive Low Pass Filter (LPF) was used to filter the sensor responses signal from unwanted noise. The passive filter was used because of its fewer components and low noise as compared to active filter. The cut off frequency was calculated by Equation 9. The LPF was designed to allow frequencies below 2 KHz to pass, while blocking frequencies above it.

$$fc = \frac{1}{2\pi RC} \quad (9)$$

Where,

- *fc* is the cut off frequency
- R is the resistor value
- C is the capacitor value

Two signal conditioning circuit boards were designed and developed for sensors located at the left and right of the sensor chamber. The circuit boards were designed using Power Logic software. Then it was converted into a schematic diagram for the board etching process by using Power PCB software. Then both signal conditioning board is etching for components placement through soldering process. The developed sensing module includes sensor chamber and signal conditioning board. The left side board is for carbon monoxide and methane while oxygen and hydrogen sulphide sensors are on the right.

4.4. Controller Board

The controller Printed Circuit Board (PCB) was also designed using Power Logic and Power PCB software. The fabrication is as shown in Figure 10. The PCB main component uses Surface Mount Technology (SMT). The electronic components were laid out on a two-layer plated through hole PCB. All the ICs were decoupled to ground with capacitors to reduce noise from the circuits. The ground plane tracks at the bottom layer were kept to a minimum to reduce the signal to Noise Ratio (SNR). The acquired data is transmitting through Zigbee wireless Radio Frequency (RF) communication.

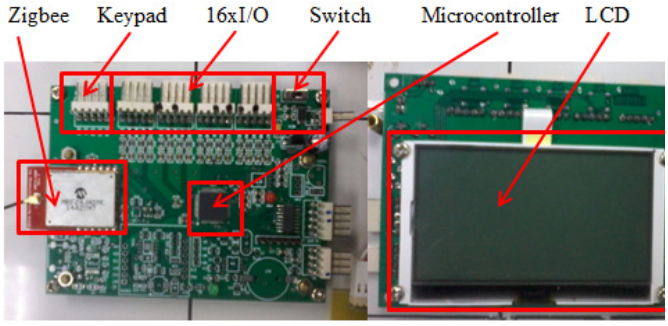


Figure 10: The controller board front and back view

4.5. Microcontroller

The e-nose system controls and data processing use the microcontroller with embedded software. The dsPIC33 microcontroller from Microchip Technology Inc. was selected because its 40 MIPS is good for high processing speed and its 16-bit data bus is suitable for the instrument operation in real time. The microprocessor flash memory is 256 KB and 30 KB for RAM are large enough and capable of handling the data acquisition operation and system control at the same time.

Furthermore, the dsPIC33's built-in 12-bit ADC conversion speed is capable of simultaneously sampling from all the analogue sensor responses signal. The digitized data is stored inside the microcontroller flash memory and sent to laptop computer which is linked through Zigbee wireless RF communication during the data acquisition process. The output digital signals are expressed in numbers ranging from 0 to 4095 and sensors voltage output is calculated by Equation 10.

$$V_{out} = \frac{ADC \times 5}{4096} \quad (10)$$

4.6. Wireless Communication

The e-nose wireless communication was developed using Zigbee (MRF24J40MC) module from Microchip Technology Inc. based on 2.4 Gigahertz (GHz) IEEE Std. 802.15.4™. This module has been chosen as a medium for communication because its communication range can reach up to 4000 feet and the temperature range is at -40 ° C to + 85 ° C which are suitable for confined space applications. It is also popular because its miniature size, simple circuit and low power. The Zigbee module has a 50Ω Ultra Miniature Coaxial (U.FL) connector to connect to an external 2.4 GHz antenna. The module interfaces to the microcontrollers through a four wire Serial Peripheral interface (SPI), interrupt, wake, reset, power and ground as shown in Figure 11.

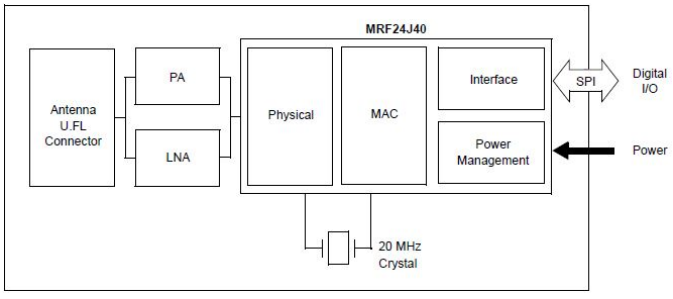


Figure 11: Wireless Zigbee (MRF24J40MC) module block diagram

4.7. Input and Output Devices

The e-nose main controller board was interfaced with a four-button keypad as the input device. The input is used as the user interface to select the instrument program menu. A 128x64 character graphic Liquid Crystal Display (LCD) from Topway Displays brand was selected as the instrument display unit (output). The LCD is used to display several instrument attributes such as operation, data acquisition and utilities.

4.8. Purging System

The sampling system will deliver the air sample to flow into the sensor chamber. The system is essential to ensure that all sensors are exposed to the air sample effectively during the sniff cycle. For this confined space application the e-nose has a self-purging function to ensure that the sensor chamber is purged accordingly [1].

Figure 12 shows a pneumatic system that has been designed as the e-nose self-purging system. The system has a 3-way solenoid electro-valve that was placed between instrument inlet and sensor chamber. During sniff cycle the electro-valve would allow the air sample from the confined space atmosphere to flow into the sensor chamber through channel one. While during purging cycle, the electro-valve would allow filtered air through the carbon filter to purge the sensor chamber cavity through channel two. The system is programmed and controlled by the instrument's microcontroller.

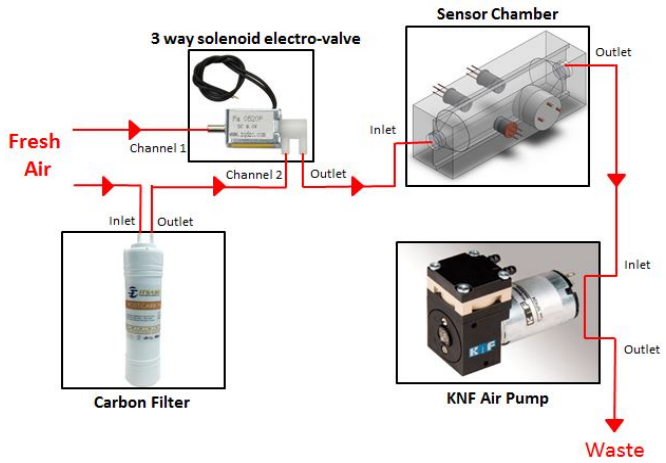


Figure 12: The e-nose purging system

4.9. Power Unit

The e-nose system uses a single 12 volt lead acid battery as the power supply. The power supply is used for the air pumps, electro-valve, signal conditioning boards and main controller board. The PCB consists of two unit voltage regulators (LM2575) used to provide 5 volts of power for the sensor measuring circuits, amplifiers, analogue multiplexer and relay.

4.10. Enclosure

The enclosure or casing for the e-nose was selected based on space for component placement and the material used. The plastic enclosure was used because of its non-flammable properties when exposed to high temperatures. It also did not interfere with data acquisition during the sniff operation. Figure 13 shows the enclosure with dimension of 22cm (L) x 30cm (W) x 14cm (H) and the instrument's complete hardware development.

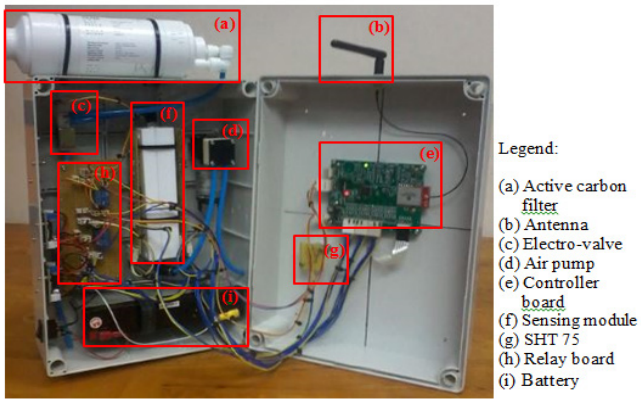


Figure 13: E-nose enclosure and hardware development

4.11. *Software Development*

The e-nose embedded software is used to integrate hardware to enable the instrument operate as a system. The acquired data transferred wirelessly from the instrument to a laptop computer that is received and displayed through Graphical User Interface (GUI). The data is processed and classified using an off-line method by utilizing MATLAB software.

4.12. *Embedded Control Software*

The e-nose embedded software has been developed on the laptop computer using dsPIC33 C compiler version 8.0 from Microchip Technology Inc. The compiler was bundled inside the MPLAB Integrated Development Environment (IDE). The menu driven program contains the instrument control, data acquisition, communication and utilities.

4.13. *Graphical User Interface*

A Graphic User Interface (GUI) was developed using Visual Basic 6.0 (VB6) software on a windows 7 platform. The GUI is used for receiving and recording the sensors output digital signal that is transmitted wirelessly by the e-nose to laptop computer. The GUI will display the sensor digital output signal that is plotted in a single bar graph according to their concentrations as shown in Figure 14. The instrument sensor digital output signals that are displayed by the GUI in real time are saved as a text file.

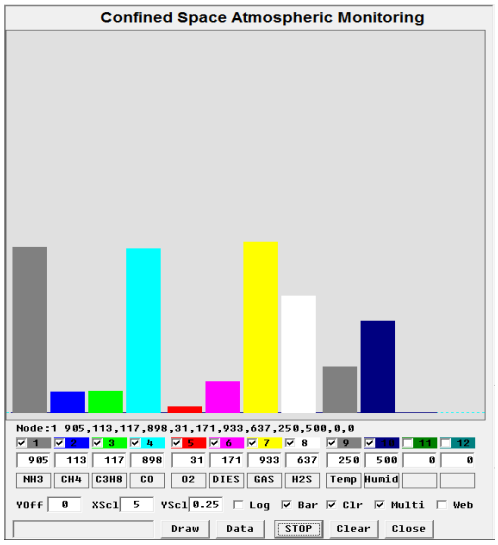


Figure 14: Graphic User Interface (GUI) for sensor responses

4.14. *Converting ADC Data to parts per million*

The developed GUI is also able to convert the e-nose sensor digital output signal as data in ADC value to parts per million (ppm) and percentage correspond to the air sample concentration. Based on the e-nose MOS gas sensors data sheet, the MOS gas sensors have a linear correlation between the sensor responses and air sample concentration. The sensor response gradient (m) and constant value (k) that intersect at $x = 1$ was calculated using Equation 11 and Equation 12.

$$m = \frac{\Delta(\log y)}{\Delta(\log x)} \tag{11}$$

$$k = \frac{y}{x^m} \tag{12}$$

Where,

- m is the gradient
- y is the sensors response (Ω) value
- x is the sample concentrations (ppm) value
- k is the intersection point at axis x=1

Based on the calculated value of the sensor response gradient and constant value, the sensor digital output signal in ppm value, x can be calculated by using Equation 13.

$$x = 10^{\left[\frac{\log(\frac{y}{k})}{m}\right]} \tag{13}$$

For the flammable gas (i.e. methane), the sensor digital output signal is measure in percentages by using Equation 14. The developed GUI for the instrument sensors digital output signal conversion is shown in Figure 15.

$$x_{\%} = \frac{x_{ppm}}{10000} \tag{14}$$

Where,

- $x_{\%}$ is the sample concentrations in percentage
- x_{ppm} is the sample concentrations in ppm

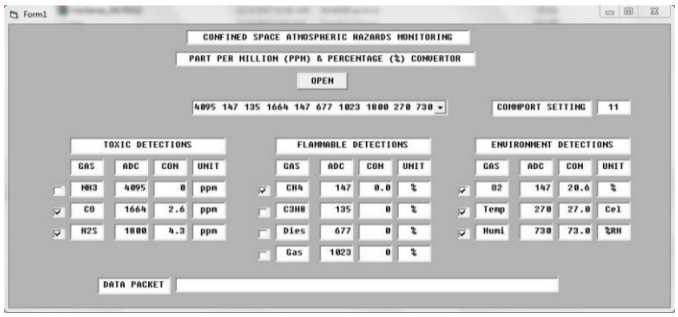


Figure 15: Graphic User Interface (GUI) for data conversion

5. Experimental Setup

The e-nose was tested in laboratory and field environments. This process will ensure the instrument functionality, performance and reliability were satisfied. The first experiment was conducted to test the performance of the purging system. After that, the instrument was calibrated to test its response to various air sample

concentrations. Lastly the instrument was tested in a hospital confined space (i.e. mechanical room) for field atmospheric hazards monitoring.

5.1. System Setup

Dynamic headspace technique was used for the e-nose sampling process. The instrument system setup was set according to the parameters listed in Table 5.

Table 5: The e-nose system setup parameter

Operation	Time (s)	Air pump	Electro-valve
Switch on e-nose and run GUI	30	OFF	OFF
Sensors pre-heat period	90	OFF	OFF
Purge cycle	60	ON	ON
Holding period after purge cycle	30	OFF	OFF
Sniff cycle	30	ON	OFF
Holding period after sniff cycle for data acquisition	200	OFF	OFF

Based on the system setup parameters, the e-nose uses the dynamic headspace sampling method includes "purge", "sniff" and "hold". The method will hold the air sample in sensor chamber to enhance the e-nose sensitivity. The setup parameter values were based on the selected sensors characteristics. The data acquisition experiment procedures are as follows:-

i. Switch on the e-nose and run the GUI on a laptop computer to test the communication.
ii. Pre-heat the instrument sensors to ensure the MOS gas sensors reach the optimum operation temperature.
iii. Activate the self-purging procedure (purge cycle) by activating the 3-way solenoid electro-valve channel two. This would allow the ambient air to flow through the active carbon filter by using the air pump. The clear reference air will flow into the sensor chamber and flush out the previous air sample. Then the instrument will remain idle for about 30 seconds to allow sensors response to return to its basic value.
iv. Sniff cycle, the 3-way solenoid electro-valve channel one is activated for the instrument expose to the air sample. The headspace air sample will flow into the sensor chamber by using the same air pump.
v. The air sample will be held in chamber for 200 seconds that will expose the sensors for optimum interaction and sensor response reach steady state.
vi. The steady state sensor response data is transfer wirelessly through the Zigbee module to a laptop computer that is displayed and saved by the GUI. The GUI will record 200 data at one data per second baud rate based on the sensor's response.
vii. Repeat steps (iii) to (vi) for another nine times that will generate 2000 data which is used for the data processing.

5.2. Purging Testing

The e-nose self-purging test experiment was conducted at the biomaterials laboratory, School of Mechatronics Engineering, Universiti Malaysia Perlis (UniMAP). The laboratory temperature was 25°C with 75% Relative Humidity. The room temperature and humidity was continually observed and recorded.

The experiment sampling process used the instrument system setup as explained in section 5.1. This experiment was conducted to test its ability to flush out the air sample from the sensor chamber enabling gas sensor responses return to its basic value. The experiment had to follow the data acquisition procedure.

Initially the instrument purge cycle operation experiments were activated using laboratory ambient air as the reference. Then the instrument purge cycle was repeated by using the self-purging system which is activated by the three-way solenoid electro-valve channel two. This allows the ambient air to flow through the active carbon filter. The reference air will flow into the sensor chamber and enables sensors response to return to their basic value. Both purging method sensor responses are analyzed for their performance.

5.3. System Calibration and Validation

E-nose calibration and validation experiments were conducted also at the same laboratory. The laboratory temperature was 25°C with 75% Relative Humidity. The room temperature and humidity was continually observed and recorded.

The experiments were conducted in laboratory in a fume hood because it will absorb any leaking of the hazardous gases as shown in Figure 16. The e-nose sampling process involves two gas sampled cylinders with different concentration as listed in Table 6. The experiment sampling process used the e-nose system setup as explained in section 5.1. The experiment had to follow the data acquisition procedure.

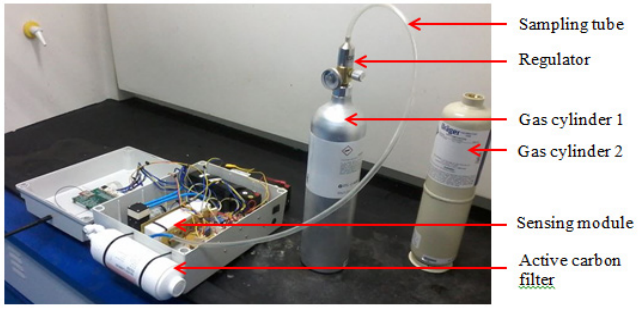


Figure 16: The e-nose calibration and validation testing

Table 6: The component and concentrations for gas cylinder one and gas cylinder two

Components	Concentrations	
	Cylinder 1	Cylinder 2
Hydrogen sulphide (H2S)	10 ppm	< 1 ppm
Carbon monoxide (CO)	50 ppm	
Methane (CH4)	2.5 %	
Oxygen (O2)	18.0 %	20.8%

The gas in cylinder one which contains oxygen, hydrogen sulphide, carbon monoxide and methane used for the calibration while cylinder two which contains air with zero grades (<1 ppm) were used for the validation.

Initially, the experiments were conducted by using gas from cylinder one (a mixing of hydrogen sulphide, carbon monoxide, methane and oxygen) as the sample to identify and calculate the calibration value. Then the experiments were repeated by using gas cylinder two which contain air with zero grades. The experiments were used to verify the e-nose ability to responses with lowest concentration of air sample.

E-nose performance was validated with an Altair 5X Multi Gas Detector from MSA brand which is able to detect hydrogen sulphide, carbon monoxide, methane and oxygen [53]. The instrument temperature and humidity measurement value were also validated by using the Humidity Alert Detector from Extech as shown in Figure 17.

(a) (b)

Figure 17: (a) Altair 5X Multi Gas Detector and (b) Humidity Alert detector

Initially both gas cylinders were used as air samples for the gas detector. The validate instrument measurement was used as the reference value. Then the experiments were conducted for the instrument using both gas cylinders. The different measurement values between the instrument and validate instrument which are known as percentage error ($\%_{error}$) by using Equation 15 were adjusted to be less than 10% by modifying the embedded instrument data acquisition program [53]. The experiments were repeated to 10 times for the instrument and validate instrument using both gas cylinders for validation process.

$$\%_{Error} = \left| \frac{\text{Experimental value} - \text{Reference value}}{\text{Reference value}} \right| \times 100 \quad (15)$$

Where,

- $\%_{Error}$ is percentage error value
- Experimental value is e-nose sensor responses data
- Reference value is validate instruments data

E-nose temperature and humidity measurement values were also compared with the commercial temperature and humidity meter. The different measurement values were also adjusted to less than 10% by modifying the embedded instrument data acquisition program. The experiments were also repeated for the instrument to validate the temperature and humidity measurements values.

5.4. Field Environment Testing

E-nose field environment experiment was conducted in a confined space (i.e. mechanical room) at the Hospital Sultanah Bahiyah, Alor Setar, Kedah. There are two set generators placed in the mechanical room to generate electricity when electricity is suddenly cut off in the hospital building. This room is potentially exposed to the atmospheric hazards when there is a leak at the generator that uses diesel fuel to operate. During the operation, the carbon monoxide gas will release and oxygen level became limit because eliminate by this gas.

This mechanical room is also chosen because of its appropriate area and the distance that allows wireless communication devices to function fully without losing any data. The room size was 20m (L) x 10m (W) x 5m (H) and the room average temperature was 31°C with 63% Relative Humidity. This room temperature and humidity condition has no effect to the sensors ability in term of drift response and sensitivity changes after calibration experiment testing were done based on the sensors operating temperature. The room temperature and humidity were continually observed and recorded.

Four locations, ambient (outside room), location one, location two and location three were selected for data acquisition. The purpose of these locations was to prove the ability of the instrument to be able to classify the concentration of gas according to their location. The experiment started with data acquisition at outside room as reference air, followed by location one, two and lastly at location three. The generator and data acquisition location in the mechanical room are illustrated in Figure 18. The experiment sampling process used the e-nose system setup as explained in section 5.1.

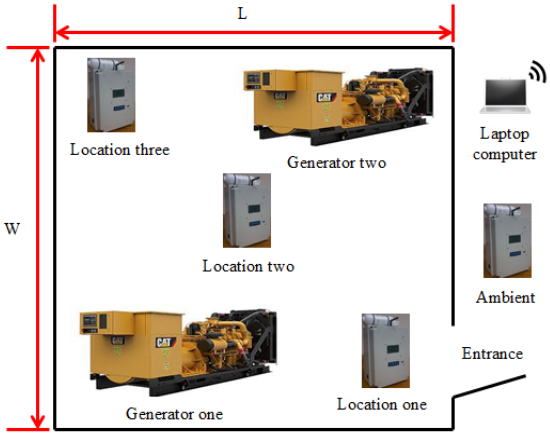


Figure 18: Data acquisition locations in mechanical room

6. Results and Discussion

6.1. Purging Testing

This section discusses the data analysis for the purging testing to compare reading between normal ambient air and the filtered ambient air. The analysis utilised the statistical multivariate technique to discriminate between two different samples.

E-nose sensor responses were a time series waveforms profile. The sensor responses purging cycle using laboratory ambient air and ambient air through an active carbon filter are shown in Figure 19 and Figure 20 respectively. The reference air flows into the sensor chamber and enables the sensor responses to return to their basic values.

The average for 10 times readings from sensor responses between the ambient air and filtered ambient air are listed in Table 7. The readings show that sensor responses for the ambient air through the active carbon filter is lower than ambient air without filter. This means that the active carbon filter is able to flush out the air sample from the sensor chamber better to enabling gas sensor responses return to its basic value and become stabilized during purge cycle.

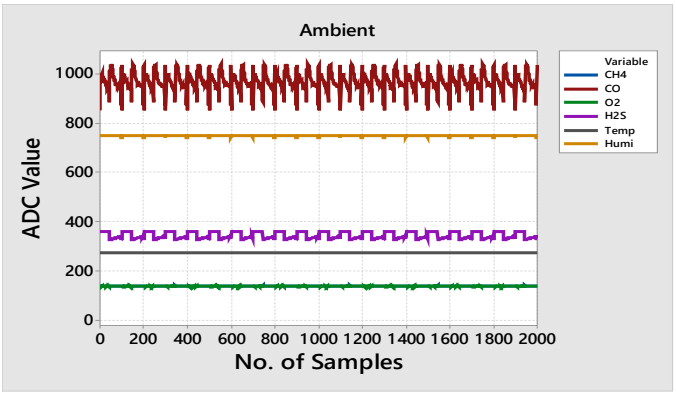


Figure 19: Sensor responses for ambient air

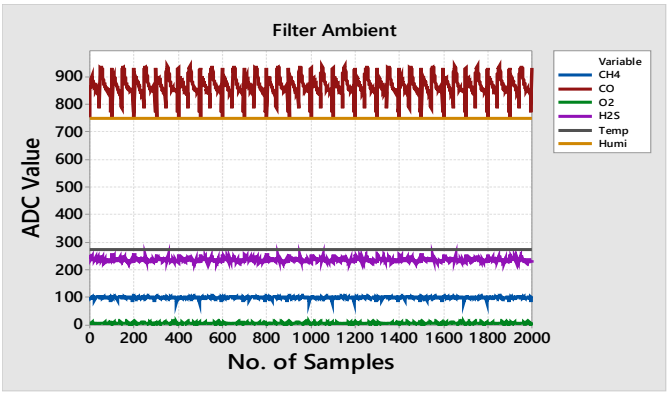


Figure 20: Sensor responses for ambient air through active carbon filter

Table 7: Average sensor responses for purging testing

Parameter	Methane (ADC value)	Carbon monoxide (ADC value)	Oxygen (ADC value)	Hydrogen sulphide (ADC value)
Ambient air	150	950	140	350
Active carbon filter	100	850	10	250

The PCA used the correlation matrix method because for all samples, the sensor responses data variances were completely different so the data variances were standardized. Figure 21 shows the PCA score plot for instrument purging testing.

Consists of six total principal component's (PC) that represent number of instrument sensors and a 2D-PCA score plot was used because the first two PC variation accounted for 99.43% (PC1 was 91.21% and PC2 was 8.22%) of the total data more than 90% which contain useful information [54]. The plot shows that data for ambient air and ambient air through an active carbon filter are successfully discriminate into two groups. This shows that the proposed self-purging system capability is successful in providing reference air for the instrument purge cycle operation.

6.2. Calibration and Validation

This section discusses the data analysis for the calibration process by calculating the calibration value. The validation between developed instrument and validated instrument is to verify its detection performance and reliability.

Sensor responses for the e-nose calibration which used gas cylinder one and gas cylinder two are shown in Figure 22 and Figure 23. The plotted graph shows that the different detection reading between both gas cylinders. The feature selection process was applied to the raw data and 100 steady state responses were select for both gas cylinders detection.

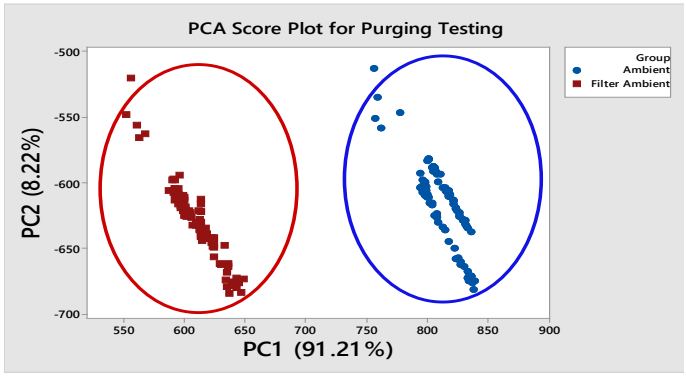


Figure 21: PCA score plot for purging testing

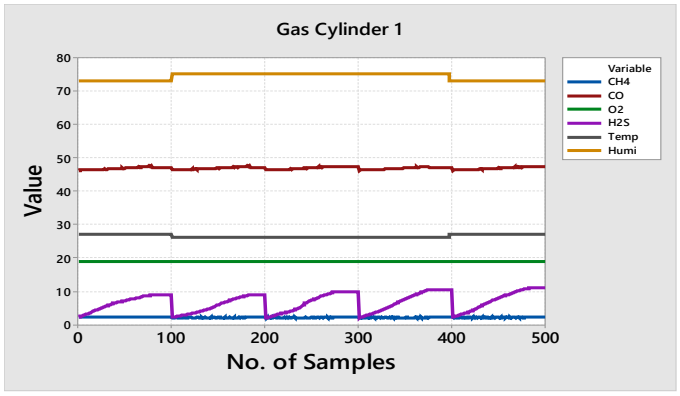


Figure 22: Sensor responses for gas exposure from cylinder one

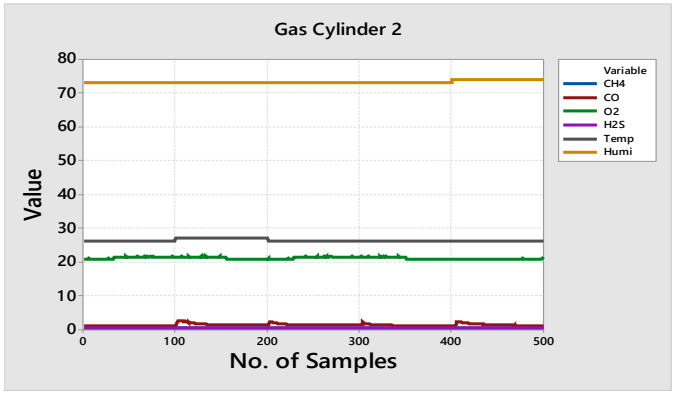


Figure 23: Sensor responses for gas exposure from cylinder two

The calibration value is the different reading between the e-nose average sniff and the commercial detector as a validated instrument. The percentage error ($\%_{Error}$) calculation for instrument sensors when exposed to the concentrations of gas cylinder one is shown in Table 8. The instrument sniff reading was 2.03% for methane, 46.83 ppm for carbon monoxide, 18.68% for oxygen, 6.68 ppm for hydrogen sulphide, 26.40°C for temperature and 74.20% Relative Humidity for humidity sensors. The different values between instrument sniff and validate instrument shown by the methane is 0.47%, carbon monoxide is 3.17 ppm, oxygen is 0.12%, hydrogen sulphide is 3.32 ppm, temperature is 0.40°C and humidity is 0.20% Relative Humidity.

Results from these different values are used as the calibration values for each sensor when performing the next detections. The percentage error was than calculated and it shows 18.80% for methane, 6.34% for carbon monoxide, 0.63% for oxygen, 33.20% for hydrogen sulphide, 1.53% for temperature and 0.27% for humidity sensors.

Table 8: Calibration value and percentage error calculation before calibration

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
E-nose sniff	2.03 ±0.08	46.83 ±0.12	18.68 ±0.02	6.68 ±0.93	26.40 ±0.36	74.20 ±0.24
Altair 5x	2.50	50.00	18.80	10.00	-	-
Humidity Alert	-	-	-	-	26.00	74.00
Calibration value	**0.47**	**3.17**	**0.12**	**3.32**	**0.40**	**0.20**
Percentage error (%)	**18.80**	**6.34**	**0.63**	**33.20**	**1.53**	**0.27**

The same sampling method process is performed, but this time by adding the calibration value for the instrument sensors when exposing them to the concentrations of gas in cylinder one. The result is shown in Table 9. The different readings between instrument sniff and validate instrument shown by the methane is 0.04%, carbon monoxide is 1.18 ppm, oxygen is 0.24%, hydrogen sulphide is 0.50 ppm, temperature is 0.60°C and humidity is 0.40% Relative Humidity. The percentage error was then calculated and it shows 1.60% for methane, 0.36% for carbon monoxide, 0.21% for oxygen, 5.00% for hydrogen sulphide, 2.22% for temperature and 0.80% for humidity sensors. Since the percentage error is decreased to less than 5.00% which is within the standard total difference allowed (less than 10%), the results were considered as valid and acceptable [55]. This proved that the instrument is trusted and ready for real field environment testing.

Table 9: Calibration value and percentage error calculation after calibration

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
Average, sniff	2.46 ±0.09	49.82 ±0.11	18.76 ±0.03	9.50 ±0.86	26.40 ±0.30	75.60 ±0.21
Altair 5x	2.50	50.00	18.80	10.00	-	-
Humidity Alert	-	-	-	-	27.00	75.00
Calibration value	**0.04**	**1.18**	**0.24**	**0.50**	**0.60**	**0.40**
Percentage error (%)	**1.60**	**0.36**	**0.21**	**5.00**	**2.22**	**0.80**

Table 10: Results of e-nose sensors ability in lowest responses

	Methane (%)	Carbon Monoxide (ppm)	Oxygen (%)	Hydrogen Sulphide (ppm)	Temperature (°C)	Humidity (%RH)
Average, sniff	**0.20** ±0.04	**0.76** ±0.43	20.94 ±0.09	**0.30** ±0.07	26.40 ±0.13	73.20 ±0.16
Altair 5x	0.00	0.00	20.80	0.00	-	-
Humidity Alert II	-	-	-	-	26.00	73.00

Table 10 shows the average lowest sniff readings for the instrument sensors during exposure to gas cylinder two. The lowest reading shown by the methane sensor is 0.20%, carbon monoxide sensor is 0.76 ppm and hydrogen sulphide sensor is 0.30 ppm. The commercial gas detector readings are also recorded to validate and to prove that the concentrations from the gas sample that have been used are trusted and reliable. The results proved that the instrument sensors are able and functioning well to response less than one ppm when exposed to the air with zero grades.

6.3. Field Environment Testing

This section discusses the data analysis for the field environment testing to test the instrument performance in real confined space. The analysis utilised the statistical multivariate and the ANN techniques.

The sensor responses for e-nose field environment testing at location one, two, three and four. Figure 24 shows the instrument sensor responses for location one air sample which consists of a dynamic slope response (transient) and steady state response. The data is pre-processed to enhance its suitability for the analysis. Feature selection is used to extract 100 steady state sensor responses in a 100 second sampling rate for pattern recognition purpose. Figure 25 shows instrument sensor responses after the feature selection process. The same feature selection process applied to other samples which are location two, three and four.

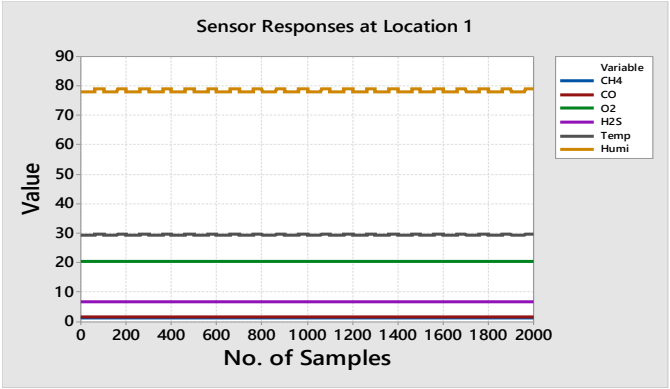


Figure 24: Location one sensor responses in field environment testing

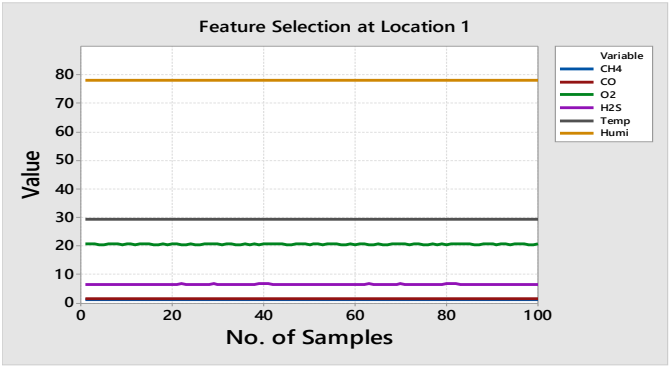


Figure 25: Location one feature selection in field environment testing

Differential baseline manipulation method was used for the air sample data. The baseline data is the different between the reference air and sample data. The process purpose is to remove the outliers and compensate the disturbance effect. This will enhance the data quality and improve sensitivity. Figure 26 shows the instrument baseline manipulation for the location one air sample.

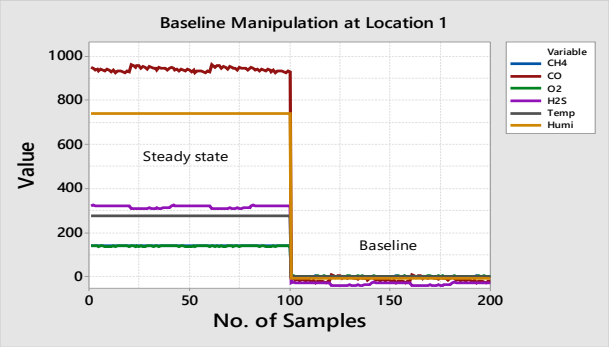


Figure 26: Location one baseline manipulation in field environment testing

A normality test was used to investigate the acquired data distribution pattern. The tests used Anderson-Darling (AD), Shapiro-Wilk (SW) and Kolmogorov-Smirnov (KS) methods. The hypothesis for the normality tests was set as:

Null hypothesis, H_0: Data is normal

Alternative hypothesis, H_1: Data is not normal

The significance level (alpha, α) for AD, SW and KS method was set at 0.05. The results shown in Table 11 indicate that the probability value (*p-value*) was less than α value for all three normality test methods. As the *p-value* is less than α, so it will reject the H_0. Thus, the acquired data distribution is not normal and the appropriate method chosen for this data is non-parametric classification analysis.

Table 11: Normality test for AD, SW and KS in field environment testing

	AD	SW	KS
Ambient	<0.005	<0.010	<0.010
Location one	<0.005	<0.010	<0.010
Location two	<0.005	<0.010	<0.010
Location three	<0.005	<0.010	<0.010

The PCA score plot for field environment testing was shown in Figure 27. Consists of six total PC's that represent number of instrument sensors and a 2D PCA score plot was used because the first two PC's variables was 99.42% (PC1 was 97.01% and PC2 was 2.41%) which contain most of the useful information 54. The plot shows that the sample data were clustered into four groups: ambient, location one, location two and location three. The plot indicates the instrument field environment testing's ability to differentiate air samples at different locations.

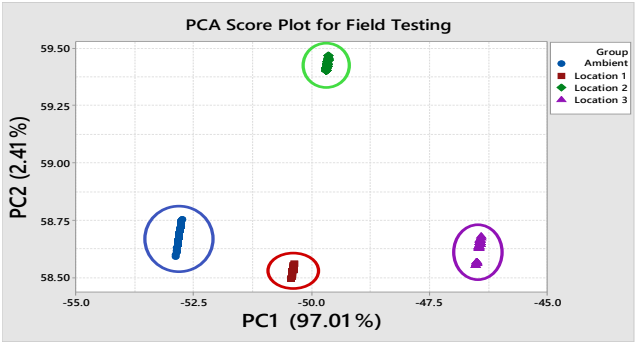


Figure 27: PCA score plot for field environment testing

The SVM is a statistical multivariate analysis method that used to test the developed instrument's capability. The methods robust features are suitable for analysing the sample data. The analysis was conducted to differentiate and classify the samples between ambient (Amb), location one (L1), location two (L2) and location three (L3).

The SVM method used 400 sample data inputs, with 100 acquired from each of the four samples. The train class was set at 70% (280 data) while test class was set at 30% (120 data). The SVM utilised the linear kernel function with prediction speed at 9400 obs/sec, training time at 2.88 sec, automatic kernel scale, box constraint at level one and the multiclass method was one versus one.

The instrument in field environment test classification success rate is shown by the confusion matrix with 99.28% for train performance and 98.33% for test performance as shown in Table 12. The results indicate that the instrument is successfully being developed and has a high classification success rate in classifying the samples.

Table 12: SVM confusion matrix for e-nose field environment testing

Train Performance						Test Performance					
		Train Class						Target Class			
		Amb	L1	L2	L3			Amb	L1	L2	L3
Class	Amb	69	1	0	0	Class	Amb	29	1	0	0
	L1	1	69	0	0		L1	1	29	0	0
	L2	0	0	70	0		L2	0	0	30	0
	L3	0	0	0	70		L3	0	0	0	30
Success rate		99.28%				Success rate		98.33%			

Table 13: RBFNN confusion matrix for e-nose field environment testing

Train Performance						Test Performance					
		Train Class						Target Class			
		L1	L2	L3	L3			Amb	L1	L2	L3
Class	Amb	69	1	0	0	Class	Amb	29	1	0	0
	L1	1	69	0	0		L2	1	29	0	0
	L2	0	0	70	0		L3	0	0	30	0
	L3	0	0	0	70		L4	0	0	0	30
Success rate		99.28%				Success rate		98.33%			

The RBFNN analysis used 400 sample data inputs, with 100 acquired from each of the four samples. The network used 70% (280 data) of the data for training and the remaining 30% (120 data) for testing. The three-layer feed-forward classification model used six input nodes which corresponds to the instrument sensors includes four MOS sensors, temperature and humidity sensor. The optimum hidden layer neuron also is set to four by using *trial* and *error* method. Four neurons were used as the output in sequences of 0001, 0010, 0100 and 1000 to help process the input which implemented the full classifications.

Table 13 shows the RBFNN classification confusion matrix for the instruments filed environment testing. The instrument classification success rate for classifying four different samples is 99.28% for train performance and 98.33% for test performance. The RBFNN classification result indicates that the instrument is successfully developed and functions accordingly.

7. Conclusion

This work has successfully developed portable e-nose for confined space application. The statistical multivariate analysis and ANN were used to model the air sample in providing potential solutions for the specific confined space application. The major achievement and contributions are summarised.

This work has successfully investigated and identified the atmospheric hazards main hazardous gases (i.e. oxygen, carbon monoxide, hydrogen sulphide and methane) that could contribute to confined space atmospheric accidents. The e-nose sensing module development was based on the main hazardous gases for confined space application.

The e-nose system was successfully fabricated with multimodal sensor detection. The instrument was able to function effectively in providing sample assessment. The instrument's key features are its portability, ease of operation and effectiveness for use in confined space application.

The e-nose system was successfully integrated with optimum self-purging. The instrument's self-purging system is able to supply the lowest sensors responses with filtered ambient air for the instrument purging cycle which is important for repeatability.

The e-nose was successfully tested its functionality in the laboratory environment through calibration and validation process using validated instruments includes Altair 5x Multi Gas Detector for concentrations of gases and Humidity Alert meter for temperature and humidity. The 5% percentage error readings between instrument and validated instruments proved its capabilities.

The e-nose was successfully tested in a real confined space environment (i.e. mechanical room) at Hospital Sultanah Bahiyah (HSB), Alor Setar, Kedah. The instrument is capable of recognising the concentrations of hazardous gasses at specific locations by discriminated using PCA with a total variation of 99.42%. While the classifiers success rate for SVM and RBFNN indicates of 99.28% for train performance and 98.33% for test performance. The instrument is able to display the atmospheric hazards concentrations in the room which is important during pre-entry testing.

Finally, based on the achievement of this research work, it is clear that the development of e-nose is able to compete existing devices to accurately and consistently measure atmospheric hazards in confine space applications.

8. Future Work

In order to enhance the instrument capabilities, a number of improvements need to be considered for future work. Firstly, improve the data conversion programming from GUI to embedded system. Secondly, improve the wireless communication distance suitable for a confined space that has a large area. It will ensure a smooth data acquisition process and prevent any data loss.

Thirdly, integrate the acquired data to send into an Internet of Things (IoT) system for real-time remote monitoring. Finally, integrate the instrument with a mobile robot for ease manoeuvring to various locations in the confined space [56]. This can eliminate an authorised person from entering into the confined space during pre-entry atmospheric testing. The mobile robot should have a complete system with a remote control and charge-coupled device (CCD) camera for any obstacle avoidance.

Acknowledgement

The author would like to acknowledge the support from the Sciencefund under grant number of 03-01-15-SF0257 from the Ministry of Science, Technology & Innovation Malaysia (MOSTI). The author also gratefully acknowledge to Universiti Malaysia Perlis (UniMAP) for the opportunities given to do the research.

References

[1] M. A. A. Bakar, A. H. Abdullah, F. S. A. Saad, S. A. A. Shukor, M. S. Kamis, A. A. A. Razak, M. H. Mustafa, "Electronic Nose Purging Technique for Confined Space Application" in IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), Penang, Malaysia, 2017. DOI: 10.1109/CSPA.2017.8064948.

[2] A. Suruda, T. Pettit, G. Noonan, R. Ronk, "Deadly Rescue: The Confined Space Hazard" Journal of Hazardous Materials, vol. 36, no. 1, pp. 45–53, 1994. https://doi.org/10.1016/0304-3894(93)E0051-3.

[3] Hazards of Confined Spaces. Vancouver: WorkSafeBC, 2008.

[4] H. Ye, "Atmosphere Identifying and Testing in Confined Space" in First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Beijing, China, 2011. DOI: 10.1109.

[5] N. B. A. Wahab, "Fatal Accident Case" Website Department of Occupational Safety and Health Malaysia, Available: http://www.dosh.gov.my/index.php/en/component/content/article/352-osh-info/accident-case/955-accident-case.

[6] T. Suski, "Common Mistakes in Confined Space Monitoring" EHS Today, Available: https://www.ehstoday.com/safety/confinedspaces/ehs_imp_37605.

[7] M. Mamat, S. A. Samad, "The Design and Testing of an Electronic Nose Prototype for Classification Problem" in International Conference on Computer Applications and Industrial Electronics, Kuala Lumpur, Malaysia, 2010. DOI: 10.1109/ICCAIE.2010.5735108.

[8] W. Chansongkram, N. Nimsuk, "Development of a Wireless Electronic Nose Capable of Measuring Odors Both in Open and Closed Systems" Procedia Computer Science, vol. 86, pp. 192–195, 2016. DOI: 10.1016/j.procs.2016.05.060.

[9] F. Röck, N. Barsan, U. Weimar, "Electronic Nose: Current Status and Future Trends" Chemical Reviews, vol. 108, no. 2, pp. 705–725, 2008. DOI: 10.1021/cr068121q.

[10] A. Wilson, M. Baietto, "Applications and Advances in Electronic-Nose Technologies" Sensors, vol. 9, no. 7, pp. 5099–5148, 2009. DOI: 10.3390/s90705099.

[11] RAE Systems by Honeywell, "Application Note 206 Guide to Atmospheric Testing in Confined Spaces" Available: https://www.raesystems.com/sites/default/files/content/resources/Application-Note-206_Guide-To-Atmospheric-Testing-In-Confined-Spaces_04-06.pdf.

[12] L. Capelli, S. Sironi, R. D. Rosso, "Electronic Noses for Environmental Monitoring Applications" Sensors, vol. 14, no. 11, pp. 19979–20007, 2014. DOI: 10.3390/s141119979.

[13] Y. Zou, H. Wan, X. Zhang, D. Ha, P. Wang, Bioinspired Smell and Taste Sensors, Dordrecht: Springer, 2015.

[14] L. Capelli, S. Sironi, R. D. Rosso, "Odor Sampling: Techniques and Strategies for the Estimation of Odor Emission Rates from Different Source Types" Sensors, vol. 13, no. 1, pp. 938–955, 2013. DOI: 10.3390/s130100938.

[15] N. E. Barbri, E. Llobet, N. E. Bari, X. Correig, B. Bouchikhi, "Application of a Portable Electronic Nose System to Assess the Freshness of Moroccan Sardines" Materials Science and Engineering: C, vol. 28, no. 5-6, pp. 666–670, 2008. DOI: 10.1016/j.msec.2007.10.056.

[16] J. Gutiérrez, M. Horrillo, "Advances in Artificial Olfaction: Sensors and Applications," Talanta, vol. 124, pp. 95–105, 2014. http://dx.doi.org/10.1016/j.talanta.2014.02.016.
[17] T. Pogfay, N. Watthanawisuth, W. Pimpao, A. Wisitsoraat, S. Mogpraneet, T. Lomas, M. Sangworasil, A. Tuantranont, "Development of Wireless Electronic Nose for Environment Quality Classification" in International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON2010), Chiang Mai, Thailand, 2010.
[18] S. Panigrahi, S. Balasubramanian, H. Gu, C. Logue, M. Marchello, "Design and Development of a Metal Oxide Based Electronic Nose for Spoilage Classification of Beef" Sensors and Actuators B: Chemical, vol. 119, no. 1, pp. 2–14, 2006. DOI:10.1016/j.snb.2005.03.120.
[19] F. Difrancesco, M. Falcitelli, L. Marano, G. Pioggia, "A Radially Symmetric Measurement Chamber for Electronic Noses" Sensors and Actuators B: Chemical, vol. 105, no. 2, pp. 295–303, 2005. DOI:10.1016/j.snb.2004.06.013.
[20] A. H. Abdullah, A. H. Adom, A. Y. Shakaff, M. N. Ahmad, M. A. Saad, "Odour Sensor Chamber Development for Electronic Nose" in 4th International Workshop on Computer Science and Engineering-Winter (WCSE), Dubai, UAE, 2014.
[21] S. M. Scott, D. James, Z. Ali, W. T. Ohare, "Optimising of the Sensing Chamber of an Array of a Volatile Detection System: Fluid Dynamic Simulation" Journal of Thermal Analysis and Calorimetry, vol. 76, no. 2, pp. 693–708, 2004. DOI: 10.1023/B:JTAN.0000034891.68585.4a.
[22] D. James, S. M. Scoot, Z. Ali, W. T. O'Hare, "Chemical Sensors for Electronic Nose Systems" Microchima Acta, vol. 149, pp. 1–17, 2005. DOI: 10.1007/s00604-004-0291-6.
[23] G. H. Jain, "MOS Gas Sensors: What Determines Our Choice?" in Fifth International Conference on Sensing Technology (ICST), Palmerston North, New Zealand, 2011. DOI: 10.1109/ICSensT.2011.6137067.
[24] M. Nissfolk, Development of an Electronic Nose-Tongue Data Acquisition System using a Microcontroller, Uppsala University, 2009.
[25] E. Noorsal, O. Sidek, J. M. Saleh, "Automated Odour Measurement in Electronic Nose System Using Microcontroller" in International Conference on Man-Machine Systems 2006, Langkawi, Malaysia, 2006.
[26] A. H. Abdullah, A. H. Adom, A. Y. M. Shakaff, M. N. Ahmad, A. Zakaria, N. A. Fikri, O. Omar, "An Electronic Nose System for Aromatic Rice Classification" Sensor Letters, vol. 9, no. 2, pp. 850–855, 2011. DOI:10.1166/sl.2011.1629.
[27] K.-T. Tang, S.-W. Chiu, M.-F. Chang, C.-C. Hsieh, J.-M. Shyu, "A Low-Power Electronic Nose Signal-Processing Chip for a Portable Artificial Olfaction System" IEEE Transactions on Biomedical Circuits and Systems, vol. 5, no. 4, pp. 380–390, 2011. DOI: 10.1109/TBCAS.2011.2116786.
[28] S. Samadi, "Interface Design Techniques for Electronic Nose Sensors: A Survey" in Sixth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, Venice, Italy, 2013.
[29] K. M. Pierce, B. Kehimkar, L. C. Marney, J. C. Hoggard, R. E. Synovec, "Review of Chemometric Analysis Techniques for Comprehensive Two Dimensional Separations Data" Journal of Chromatography A, vol. 1255, pp. 3–11, 2012. http://dx.doi.org/10.1016/j.chroma.2012.05.050.
[30] L. A. Berrueta, R. M. Alonso-Salces, K. Héberger, "Supervised Pattern Recognition in Food Analysis" Journal of Chromatography A, vol. 1158, no. 1-2, pp. 196–214, 2007. DOI:10.1016/j.chroma.2007.05.024.
[31] J. Yan, X. Guo, S. Duan, P. Jia, L. Wang, C. Peng, S. Zhang, "Electronic Nose Feature Extraction Methods: A Review" Sensors, vol. 15, no. 11, pp. 27804–27831, 2015. DOI:10.3390/s151127804.
[32] J. Lozano, J. P. Santos, M. C. Horrillo, "Enrichment Sampling Methods for Wine Discrimination with Gas Sensors" Journal of Food Composition and Analysis, vol. 21, no. 8, pp. 716-723, 2008. DOI:10.1016/j.jfca.2008.07.002.
[33] R. Sheebha, M. Subadra, "A Survey on Various Signal Pre-Processing Methods for Sensor Response in E-Nose" Researcher, vol. 6, no. 11, pp. 76–80, 2014. http://www.sciencepub.net/researcher.
[34] N. M. Razali, Y. B. Wah, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests" Journal of Statistical Modeling and Analytics, vol. 2, no. 1, pp. 21–33, 2011.
[35] N. A. Ahad, T. S. Yin, A. R. Othman, C. R. Yaacob, "Sensitivity of Normality Tests to Non-normal Data" Sains Malaysiana, vol. 40, no. 6, pp. 637–641, 2011.
[36] A. Ghasemi, S. Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non Statisticians" International Journal of Endocrinology and Metabolism, vol. 10, no. 2, pp. 486–489, 2012. DOI: 10.5812/ijem.3505.
[37] L. I. Smith, A Tutorial on Principal Components Analysis Introduction, 2002.
[38] C. Cortes, V. Vapnik, "Support-Vector Networks" Machine Learning, vol. 20, no. 3, pp. 273-297, Kluwer Academic Publishers, 1995.
[39] Z. Haddi, A. Amari, H. Alami, N. El Bari, E. Llobet, B. Bouchikhi, "A Portable Electronic Nose System for the Identification of Cannabis-based Drugs" Sensors and Actuators B: Chemical, vol. 155, no. 2, pp. 456-463, 2011. DOI: 10.1016/j.snb.2010.12.047.
[40] L. Auria, R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis" SSRN Electronic Journal, 2008. DOI: 10.2139/ssrn.1424949.
[41] I.-C. Chen, J. K. Hill, R. Ohlemuller, D. B. Roy, C. D. Thomas, "Rapid Range Shifts of Species Associated with High Levels of Climate Warming" Science, vol. 333, no. 6045, pp. 1024–1026, 2011. DOI: http://ssrn.com/abstract=1424949.
[42] R. Domínguez, L. Moreno-Barón, R. Muñoz, J. Gutiérrez, "Voltammetric Electronic Tongue and Support Vector Machines for Identification of Selected Features in Mexican Coffee" Sensors, vol. 14, no. 9, pp. 17770–17785, 2014. DOI: 10.3390/s140917770.
[43] J. Brezmes, N. Canyellas, E. Llobet, X. Vilanova, X. Correig, "Application of Artificial Neural Networks to the Design and Implementation of Electronic Olfactory Systems" in 5th Seminar on Neural Network Applications in Electrical Engineering (NEUREL 2000), Belgrade, Yugoslavia, 2000. DOI: 10.1088/0957-0233/1/5/012.
[44] T. Li, "The Application of Artificial Neural Networks in the Electronic Nose for Odour Measurement," M.Sc. Thesis, University of Manitoba, 2005.
[45] K. Suzuki, Artificial Neural Networks - Industrial and Control Engineering Applications, Rijeka: InTech, 2011.
[46] C. S. Kudarihal, M. Gupta, "Electronic Nose Based on Metal Oxide Semiconductor Sensors as an Alternative Technique for Perception of Odours" International Journal of Advances in Engineering & Technology, vol. 7, no. 1, pp. 206–216, 2014.
[47] M. Ghasemi-Varnamkhasti, S. S. Mohtasebi, M. Siadat, S. Balasubramanian, "Meat Quality Assessment by Electronic Nose (Machine Olfaction Technology)" Sensors, vol. 9, no. 8, pp. 6058–6083, 2009. DOI: 10.3390/s90806058.
[48] B. Tudu, A. Jana, A. Metla, D. Ghosh, N. Bhattacharyya, R. Bandyopadhyay, "Electronic Nose for Black Tea Quality Evaluation by an Incremental RBF Network" Sensors and Actuators B: Chemical, vol. 138, no. 1, pp. 90–95, 2009. DOI: 10.1016/j.snb.2009.02.025.
[49] H. D. Leon-Delgado, R. J. Praga-Alejo, D. S. Gonzalez-Gonzalez, M. Cantú-Sifuentes, "Multivariate Statistical Inference in a Radial Basis Function Neural Network" Expert Systems with Applications, vol. 93, pp. 313–321, 2018. DOI: 10.1016/j.eswa.2017.10.024.
[50] N. P. Thanh, Y.-S. Kung, S.-C. Chen, H.-H. Chou, "Digital Hardware Implementation of a Radial Basis Function Neural Network" Computers & Electrical Engineering, vol. 53, pp. 106–121, 2016. http://dx.doi.org/10.1016/j.compeleceng.2015.11.017.
[51] K. Wetchakun, T. Samerjai, N. Tamaekong, C. Liewhiran, C. Siriwong, V. Kruefu, A. Wisitsoraat, A. Tuantranont, S. Phanichphant, "Semiconducting Metal Oxides as Sensors for Environmentally Hazardous Gases" Sensors and Actuators B: Chemical, vol. 160, no. 1, pp. 580–591, 2011. DOI: 10.1016/j.snb.2011.08.032.
[52] M. A. A. Bakar, A. H. Abdullah, F. S. A. Sa'Ad, S. A. A. Shukor, M. H. Mustafa, M. S. Kamis, A. A. A. Razak, S. A. Saad, "Electronic Nose Sensing Chamber Design for Confined Space Atmospheric Monitoring" in International Conference on Mathematics, Engineering and Industrial Applications 2016 (ICoMEIA), Songkhla, Thailand, 2016. http://dx.doi.org/10.1063/1.4965179.
[53] M.A.A. Bakar, A.H. Abdullah, F.S.A. Sa'ad, S.A.A. Shukor, A.A.A. Razak, M.H. Mustafa, "Electronic Nose Calibration Process for Monitoring Atmospheric Hazards in Confined Space Applications" Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2016.
[54] A. F. Najmuddin, I. M. Ibrahim, S. R. Ismail, "A Simulation Approach: Improving Patient Waiting Time for Multiphase Patient Flow of Obstetrics and Gynecology Department (O&G Department) in Local Specialist Centre" WSEAS TRANSACTIONS on MATHEMATICS, vol. 9, no. 10, pp. 778–790, 2010.
[55] K.-T. Tang, S.-W. Chiu, C.-H. Pan, H.-Y. Hsieh, Y.-S. Liang, S.-C. Liu, "Development of a Portable Electronic Nose System for the Detection and Classification of Fruity Odors" Sensors, vol. 10, no. 10, pp. 9179–9193, 2010. DOI: 10.3390/s101009179.
[56] H. Ishida, Y. Wada, H. Matsukura, "Chemical Sensing in Robotic Applications: A Review" IEEE Sensors Journal, vol. 12, no. 11, pp. 3163–3173, 2012. DOI: 10.1109/JSEN.2012.2208740.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 213-221 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

iSensA - A System for Collecting and Integrating Sensor Data

João Manuel Leitão Pires Caldeira[1], Vasco Nuno da Gama de Jesus Soares[1,*], Pedro Miguel de Figueiredo Dinis Oliveira Gaspar[2], Joel José Puga Coelho Rodrigues[3], Ricardo Manuel Valentim Fontes[4], José Luís Lopes Silva[4]

[1]*Instituto Politécnico de Castelo Branco, Instituto de Telecomunicações, Portugal*

[2]*Universidade da Beira Interior, Portugal*

[3]*National Institute of Telecommunications (Inatel), Brazil; Instituto de Telecomunicações, Portugal; University of Fortaleza (UNIFOR), Brazil*

[4]*Instituto Politécnico de Castelo Branco, Portugal*

ARTICLE INFO

Article history:
Received: 28 September, 2018
Accepted: 26 October, 2018
Online: 10 November, 2018

Keywords:
iSensA
Arduino
Sensors
Monitoring system
Web and mobile applications
Internet of Things

ABSTRACT

The idea of monitoring several types of parameters in various environments has been motivating significant research works in Internet of Things (IoT). This paper presents the design and construction of iSensA, a system for integrating and collecting information from sensors. The solution implements a multi-sensor monitoring system and then expands the monitoring concept to an IoT solution, by employing multi-network access, Web services, database and web and mobile applications for user interaction. iSensA system is highly configurable, enabling several monitoring solutions with different types of sensors. Experiments have been performed on real application scenarios to validate and evaluate our proposition.

1. Introduction

This paper is an extension of work originally presented in conference CISTI 2018 [1].

Nowadays, industry faces new challenges and paradigms derived from the current industrial revolution, the 4th one, named as Industry 4.0. This new era of industrial development aims to provide an expedite answer to the fast and dynamic requirements of production, and promoting its operational effectiveness and efficiency, and thus contributing to higher productivity. As described by [2] and [3], the most relevant features of Industry 4.0 are related with information and communication technologies & electronics (ICT&E), advanced algorithms, added value and knowledge management. These features can be summarized as: (1) digitization, optimization, and production customization; (2) automation and adaptation; (3) human machine interaction (HMI); (4) value-added services and businesses, and (5) automatic data exchange and communication. To accomplish these objective, Industry 4.0 relies on several technologies, being mobile computing, cloud computing, big data, and the Internet of Things (IoT) the key technologies of Industry 4.0 [3-7]. Among these concepts, the IoT paradigm has been extensively focused in the last years, since it covers a wide scope of applications in various fields. It brings ubiquitous intelligence through the interconnection of equipment (sensors, devices, etc.) to the Internet.

Results from a market and literature survey of IoT-based monitoring show that some developments have been reported in implementation of such systems for example for monitoring: water level [8,9] and water quality [10,11], power consumption [12,13], smart agriculture [14,15], garbage [16,17], solar energy production [18], air quality [19], equipment working status [20]. Instead of focusing on a specific problem and its solution, the work presented in this paper proposes a generic IoT data acquisition solution, that can be widely applied to different scenarios. It addresses the following main challenges:

- Applicable to different environments, with distinct requirements;
- Send and receive data to and from remote servers over Ethernet, Wi-Fi or GSM/GPRS;
- Configuration of operation and control rules based on parameters and specific requirements;
- Statistical tools to support real-time or historical data management and analysis, trend analysis, and data correlation;
- Configuration of reports, alarms/notifications, which can be sent by e-mail and/or SMS;

* Vasco Soares, Email: vasco.g.soares@ipcb.pt

- High scalability;
- Online platform accessible via computer, tablet or smartphone;
- Complete and secure backup in the cloud;
- Implemented using open source technology.

The main contributions of this work are as follows. Firstly, an intelligent data acquisition system, called iSensA [31], is proposed. It is based on Arduino controllers and sensors, Web and mobile applications supported by a cloud architecture with a relational database and Web services. Secondly, the effectiveness of this system for monitoring, analyzing real-time or historical data and ensuring incident detection is tested. The proposed system is experimentally evaluated and validated in several real-world case studies of different sectors with specific requirements. Thirdly, it is assessed if the system facilitates resource management and contributes to reduce operational expenses.

The remainder of this paper is organized as follows. Section 2 introduces iSensA system for remote monitoring. Section 3 presents the test and validation of iSensA in real scenarios, and discusses resource consumption. Section 4 concludes the paper and points further research directions.

2. iSensA

iSensA system was developed using a three-tier architecture [21] that comprises a presentation tier, a business tier, and a data tier. The three layers in the three-tier architecture are as follows. The client layer contains the user interface part of the application. Data is presented to the user or input is taken from the user. The business layer acts as an interface between the client layer and the data access layer. It is responsible for business logic like validation of data, calculations, data insertion, etc. Finally, the database is located in the data layer.

Figure 1 presents the iSensA architecture. The Server #1 is responsible for data processing. It calculates indirect values such as total, average, minimum, and maximum values. These values are calculated from direct values sent by the controllers. Indirect values are used for charts and for reporting purposes. The Server #2 is also used due to the large amount of data potentially collected from the controllers. It allows dividing the load between web / mobile users and controllers. The platform architecture was conceived as a modular system suitable to be integrated into third-party middleware.

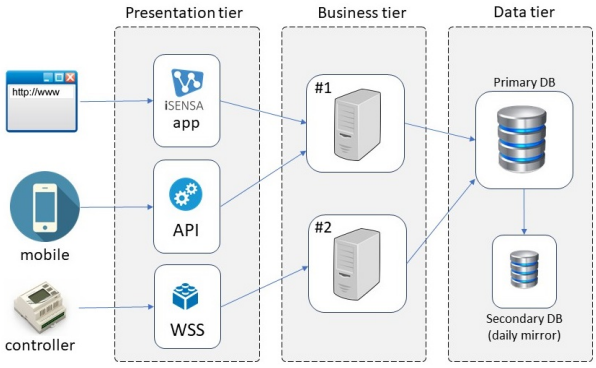


Figure 1. Overview of the iSensA system.

iSensA development required: (1) the construction of a data acquisition and control module, performed through a controller, sensors and electronic circuits; (2) the implementation of a data acquisition and processing interface, Web and mobile applications, supported by a cloud architecture with a relational database and Web services. The following subsections present the controller, the Web application, and the mobile application.

Due to space limitations it is not possible to present the features of the system and applications in detail. The interested reader may contact the authors for detailed information and request access to test it.

2.1. Controller

The iSensA controller is an Arduino-based hardware platform [22]. The adoption of this platform in development environments allows rapid prototyping of functional solutions. Characterized by the use of Atmel ARV microcontrollers, with intuitive programming based on the C programming language, this platform is becoming a true competitor to more complex control systems currently used by companies [23]. A great advantage of this type of platform is its flexibility. Providing multiple inputs and outputs, digital and analog, these platforms become adaptable to the acquisition of any type of signals. With the increasing success and the low-cost of these platforms, several sensors are available for measurement and collection of various parameters. Measuring sensors, which lately only existed for industrial controllers, are now available for use on low-cost platforms such as Arduino. Modularity is another great feature of these platforms, so they can be built according to the specific needs of a particular application.

Taking advantage of all these features, the iSensA controller is based on an Arduino platform. Depending on the requirements of each application scenario, an iSensA controller can be coupled to a communication module using Ethernet, wireless (Wi-Fi) or mobile Global System for Mobile Communications (GSM). These modules are required for communication between the controller and the iSensA Server #2. Figure 2 presents the iSensA controller in several of its different configurations.

The operation of iSensA controllers is based on collecting data from the attached sensors and sending them to the iSensA Server #2 (Figure 1). For that, the controller starts by trying to get Internet connection thru the DHCP service. After, it enters into an infinite loop divided into three main modules, "*Register Control*", "*Collect Data*" and "*Sensor Input*". The first module to be executed is "*Register Control*" and it is executed only once. This module is responsible for the registration of controllers with the iSensA Server #2. The registration procedure improves security to the system by only allowing pre-approved controllers to communicate with iSensA Server #2. Therefore, each controller before start sending data must be known by iSensA Server #2. After successful registration, the controller starts executing "*Collect Data*" and "*Sensor Input*" modules periodically. "*Collect Data*" reads the real-time values from the attached sensors while "*Sensor Input*" performs all the tasks needed to send the collected data to Server #2. Figure 3 presents the flowchart of the firmware executed by iSensA controllers.

Figure 2. iSensA controller: (a) Initial prototype of the controller with a GSM module; (b) Current version of the controller with an Ethernet module; (c) Controller with a Wi-Fi module.

All data exchanged between iSensA controllers and the Server #2 is encrypted using a proprietary algorithm to ensure data integrity and privacy. iSensA Server #2 provides two secure Web services for controllers' communication – "*WSRegisterControl*" and "*WSSensorInput*". The first one is used by "*Register Control*" module. The information sent to this Web service is the following: MAC address of the controller; Time interval for sending data; Controller model; Sensor list with identification (ID) of each sensor and type. "*WSSensorInput*" is used by "Sensor Input" module. The information sent to this Web service is the following: MAC address of the controller; Sensor list with identification of each sensor and the corresponding value. These Web services ensure that all the data collected by controllers reaches the iSensA server with the proper context. The MAC address identifies the controller that sent the data and the sensor ID the sensor to which the value corresponds.

To ensure continuous operation of the controllers some error recovery routines were included into the algorithm. If controllers cannot communicate with iSensA Server #2 within a number of attempts, then controllers reset themselves. This may occur whenever the Internet connection has been lost and must be restarted again. Besides that, all the firmware execution is shielded by a reset trigger associated to a timeout. If for some reason the firmware gets stuck at some point of its execution, beyond the timeout, then the reset trigger is activated and makes the controller to reset itself. These mechanisms guarantee robustness to the system, allowing it to recover from most uncontrolled situations.

The first tests of the system in real-world deployments revealed several limitations and challenging issues. For example, there were lots of problems concerning network communications, in particular, GSM communication. It was observed that the controller would stop working after losing GSM connection, which occurred very often. An analysis of this problem revealed that the native Arduino library procedure responsible for getting a link to the GSM network was blocking. Therefore, a timeout recovery algorithm was implemented in the GSM library to prevent the firmware to get infinitely stuck waiting for a GSM link.

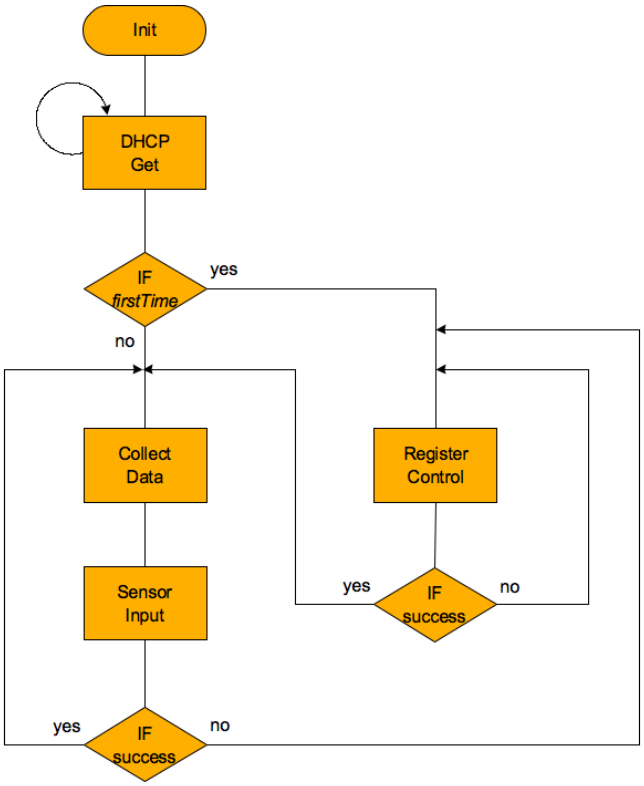


Figure 3. iSensA controller firmware flowchart.

Another issue that occurred on some iSensA real-world deployments, was caused by the signal loss in cabling between sensors and controllers. After thorough electrical analysis and several tests, it was concluded that this problem resulted from two main reasons. The first one due to long cables. This was solved by decreasing the pullup resistors used in the sensors connections, thus allowing the controllers to receive the correct values from sensors. The second one resulted from cables that were installed in noisy environments, near to pumps, electrical links, etc. In such situations, the problem was solved by using shielded cables to connect sensors to controllers.

2.2. Web Application and Relational Database

Open-source technologies were used to develop an intuitive and user-friendly Web application that can be used at any time, from anywhere, in different electronic devices (computers, tablets, mobile phones, etc.). PHP runs on the server side, being embedded in HTML. JavaScript is a scripting language executed on the client side, interpreted by the web browser, and allows the creation of interactive pages. jQuery is a quick, small and feature rich Javascript API, making HTML and event handling, animations and Ajax much simpler and easier to use. On the client side Bootstrap was also used. It is a front-end framework that allows agile development in responsive web projects. It allows integration of jQuery with other technologies such as CSS and LESS, for the dynamism desired in the interaction with the user. Bootstrap is also used in the development of platform design. Google Charts framework is used to plot data into rich and interactive charts.

Due to the characteristics of the Web application to be developed, the ICONIX methodology was selected. The analysis, design and implementation phases of the application were supported by its processes.

The Web application shown in Figure 5 provides, among others, the following functionalities:

- Secure Web access available anytime anywhere;
- Accessible through any of the major Web browsers on any device;
- Configurable dashboards;
- Easy and intuitive navigation;
- Centralized management of facilities;
- Setting access levels for different users;
- User defined alarms that can be received by email or cell phone text message;
- User defined reports that can be received by email;
- Real-time data visualization;
- Historical data analysis;
- Data export in various formats.

The database presented in Figure 6, is implemented in a MySQL fully normalized schema that allows for rapid complex querying. HTTPS was set up to protect the integrity of the Web application, and the security of its users. It allows users to securely connect to the Web application.

2.3. Mobile Application

The mobile application named MobiSensA [1] aims to extend the range of features of iSensA system to mobile devices. It allows monitoring the installations, showing the status and information of controllers and sensors in real time, without having to depend on the Web application and email. Furthermore, it notifies users when alerts or anomalies occur.

ICONIX methodology was also used for the development of the mobile application. MobiSensA was developed for smartphones with Android OS, using the Android Studio integrated development environment, and the SQLite Database Browser tool for database management. To implement the application and organize the interface, a set of specifications and guidelines of the Material Design language, referring to usability and accessibility was used. The Git tool was used to enforce versioning. Firebase technology was used to keep the application's local repository updated. A robust and secure Web Services API provides access to data from the iSensA database. Used protocols and technologies are illustrated in Figure 4.

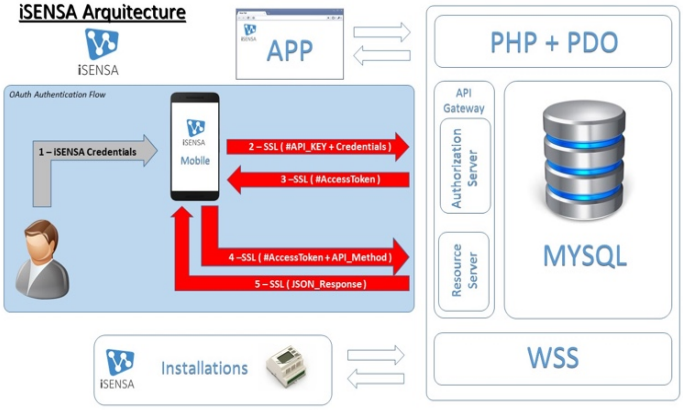


Figure 4. System's architecture.

The following are the requirements for the mobile application.

Non-functional requirements:

- The application is developed for Android 4.4 or later.
- The application limits size of stored data.
- The application works without an internet connection (offline).

Functional requirements:

- The user must login to use the application;
- The application shows the user's installations, together with the status of the controllers;
- The application allows visualizing the entire structure of the user data (installations, controllers and sensors);
- The application allows visualizing a sensor details, readings and alert history;
- The application icon must show the number of installation alerts and/or anomalies;
- The user may control the update frequency of the controllers/sensors;
- The user may decide the number of days the information is kept in history;
- The application notifies the user when an anomaly or an alert occurs;
- The application shows the number of anomalies and alerts that have occurred.

Some screenshots of the mobile application are shown in Figures 7-9. Figure 7 a) shows the start interface with a login option; b) displays the user's installations and controllers. Figure 8 a) shows detailed information of a controller; b) detailed information of a sensor that is connected to a controller. Figure 9 a) shows the sensor alerts; b) notification of controller anomalies.

Figure 5. iSensA Web Application.

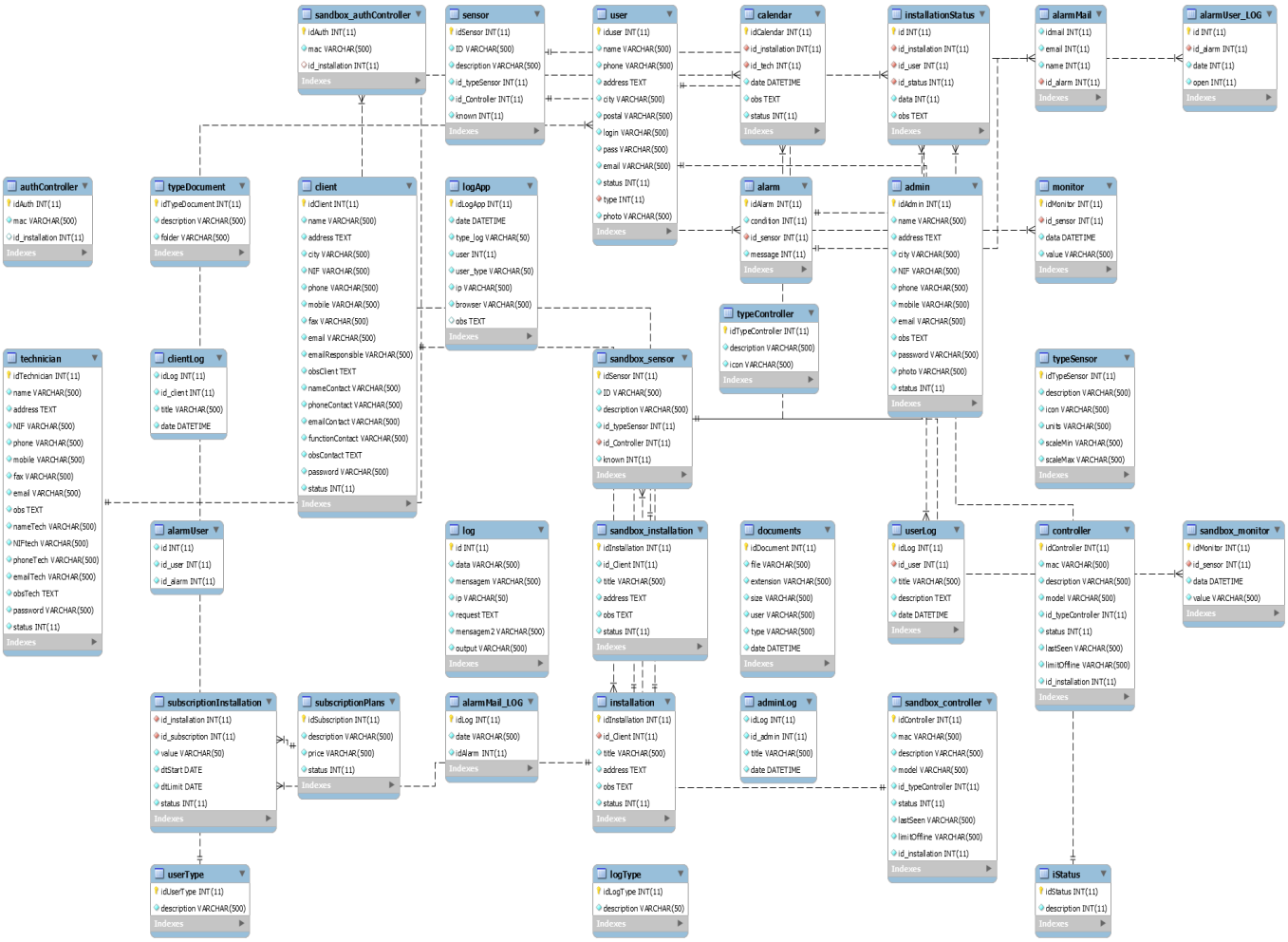
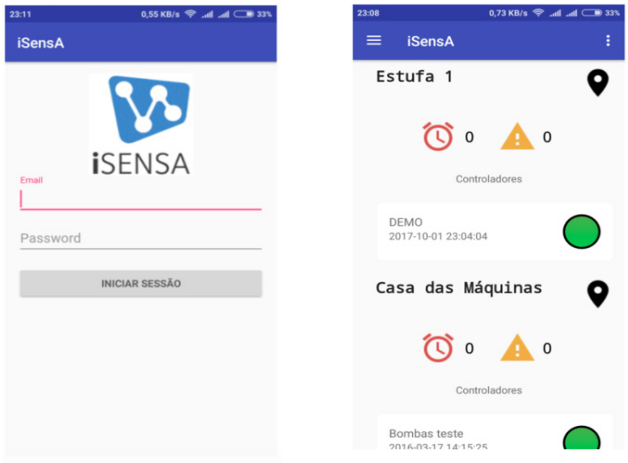


Figure 6. iSensA Relational Database.



a) b)

Figure 7. Screenshots: a) login; b) user's installations.

a) b)

Figure 8. screenshots: a) detailed information of a controller; b) detailed information of a sensor.

This section presents two real application scenarios that were used for testing and validation of the iSensA system. The evaluation of the iSensA resource consumption is also discussed.

ALBIGEC [24], a company responsible for the management of the swimming pool complex in the city of Castelo Branco, Portugal, wanted to monitor and remotely control the swimming pools water temperature, the bath waters temperature, and the female and male bathrooms temperature and humidity. iSensA's first prototype was deployed there. After a detailed study of the site and the requirements meeting, it was concluded that it would be necessary to install two iSensA controllers. One of these controllers was installed next to the swimming pool. Using three temperature-humidity sensors it monitors the temperature and humidity of the bathrooms and the central building. In order to monitor the water temperature of the two swimming pools and the sanitary waters, another controller was installed in the technical zone, next to the water tanks.

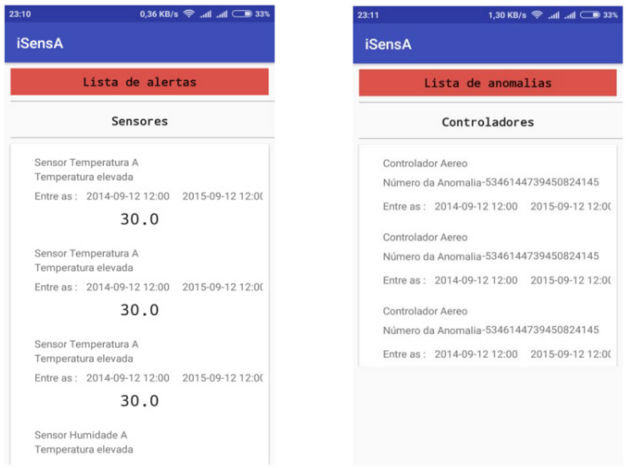


Figure 9. Screenshots: a) sensor alerts; b) controller anomalies.

(a) (b) (c)

Figure 10. Sensors for swimming pool monitoring: (a) Sensing the bath waters temperature; (b) Sensing the temperature and humidity of the environment in the main pool; (c) Sensing the temperature and humidity in a bathroom.

In this case, four temperature sensors were placed in the outlet pipes that go from the water tanks to the swimming pools, and from the sanitary water tanks to the bathrooms. Both controllers communicate with the iSensA Web services using the building's Ethernet LAN.

Following the initial success, ALBIGEC extended the use of iSensA to another swimming pool complex under its management, located in Alcains village, Portugal. This was a simpler scenario where only one controller was required. In this case, the parameters to be monitored were the temperature and humidity in two bathrooms (female and male), the temperature of the swimming pool water and the temperature of the bath waters. Since an Ethernet connection wasn't available, a GSM modem was connected to the controller to allow the communication with the iSensA Web services.

The Municipality of Pampilhosa da Serra, Portugal [25], has also successfully adopted iSensA for controlling their swimming pool complex. Figure 10 shows the installation of iSensA in this scenario. Figure 5 displays the iSensA Web platform showing the values collected in this installation.

The functionalities provided by iSensA Web application allowed both companies to program alarms with upper and lower limits for all measured parameters. These alarms trigger alerts notified by e-mail and allow companies to react accordingly. These notifications can help prevent any downtime from the swimming pool complex due, for example, to low water temperatures. iSensA also provides automated statistical reports essential to effective management. Before using iSensA, only a limited amount of data was collected on paper forms by the maintenance staff and entered into the computer at a later date. Therefore, it was difficult to sort through and analyze it in an efficient manner.

3.2. Monitoring and Control of Sewer Pumping Stations

SMAS [26], a company responsible for the municipal services in the city of Castelo Branco, Portugal, proposed as a challenge the use of iSensA to monitor a sewer pumping station. Sewer pumping stations are used to move wastewater to higher elevations in order to allow transport by gravity flow. Sewage is fed into and stored in a wet well. When the sewage level rises to a predetermined level, an electric pump is started automatically to lift the sewage upward through a pressurized pipe system from where it is discharged into a gravity manhole again. This process is repeated until the sewage reaches a treatment plant. No tool exists currently for remotely monitoring SMAS stations. Their workers visit each of the sewer pumping stations once a day, to check for problems or failures. If a fault occurs between visits, it will not be detected, and it will lead to overflows of wet wells that must be avoided due to potentially posing health and environment hazards.

iSensA was deployed in one sewer pumping station to allow for control and remote monitoring of events that occurred there. Figure 11 shows the installation in this scenario. The deployed system was comprised of an iSensA controller connected to three amperometric clamps and a level sensor. The current clamps are used to measure the instantaneous current consumption of each phase of the electric pump (in this case a three-phase electric pump). The level sensor measures the height of the water in the wet well in real time. In addition, the controller has three relays connected. When the well float drops, an overflow will occur, and the first relay is triggered. When there is a malfunction in the electric pump soft-start, the second relay is triggered. Finally, a differential protection relay can also be triggered. The second and third relays can help detecting if the pump did not start. Figure 12 presents the electric block diagram of the installed controller.

An SMS notification is sent to the telephone numbers pre-programmed on the controller, each time one of the relays is triggered. This allows workers to be notified immediately. Again the Web application allows to set alert values for both the values measured by the amperometric clamps and the water level in the well, and these alarms are then notified by e-mail. In the absence of a cabled network connection, a GPRS/GSM module was connected to the controller. This allows sending data to iSensA Web services and also sending SMS messages. SMAS workers can now react more quickly and resolve the problems sooner.

3.3. Performance and Resource Usage

The above-described scenario of the swimming pool complex of the Municipality of Pampilhosa da Serra was used for assessing iSensA system performance and resource consumption. The characterization of workload and resource consumption was performed based on server logs, including both data exchanged between iSensA controllers and the Server #2 (Figure 1) and resource consumption of the Web services requests, as well as the volume of data stored in the database.

The two iSensA controllers deployed in this scenario, acquire sensor data and send it to the server every 15 minutes. For example, during the month of October 2017, 14233 data messages were sent from the controllers to the server. The average size observed for register and data messages was about 210.18 octets. The collected sensor data over this period of time required approximately 18.06 MB of storage space in the database. Total bandwidth consumption by the Web services was 2995794 octets. These results allow to conclude about the effectiveness and efficiency of iSensA system.

4. Conclusions

It is predicted that Industry 4.0 will accelerate industry to achieve remarkable levels of operational efficiencies and thus largely increase productivity [27,28]. Among technologies, IoT is considered as a key enabler for the new-generation advanced manufacturing, Industry 4.0 [29]. IoT can be seen as a mean of connecting physical objects to the Internet as a ubiquitous network that enables objects to collect and exchange information. Thus, it blends and expands the traditional automation systems and industrial informatics systems into a much broader context [30].

In this paper, a system for reading and acquiring sensor data - iSensA - was presented. iSensA is based on Arduino controllers, web and mobile applications supported by a cloud architecture with a relational database and web services. iSensA enables the monitoring and analysis of data, environments and devices in a wide range of areas. The testing and validation of iSensA in real scenarios, show the soundness of the presented system to adapt to different application requirements. It contributes efficiently to incident detection, resource management, and cost reduction.

Future work includes: (1) using the gathered historical data to develop predictive models to help anticipate potential future events /incidents and to identify energy saving opportunities; (2)

(a) (b)

Figure 11. Monitoring a sewer pumping station: (a) Installation of the controller; (b) Wet well.

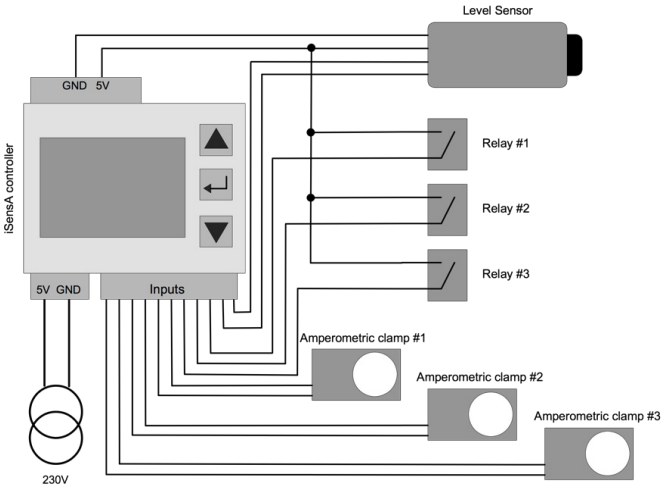


Figure 12. Electric block diagram of the installed controller.

assessing iSensA system for safety critical applications; (3) testing and maintenance of the mobile application.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER – PT2020 partnership agreement under the project UID/EEA/50008/2013, and by Finep, with resources from Funttel, Grant No. 01.14.0231.00, under the Centro de Referência em Radiocomunicações - CRR project of the Instituto Nacional de Telecomunicações (Inatel), Brazil. The authors would like to acknowledge the companies InspiringSci, Lda, Enérgico Balanço Unipessoal Lda, ALBIGEC - Empresa de Gestão de Equipamentos Culturais, Desportivos e de Lazer, E. M., and SMAS - Serviços Municipalizados de Castelo Branco, as well as the Municipality of Pampilhosa da Serra, for their interest and valuable contribution to the successful development of the iSensA system.

References

[1] J. Silva, R. Fontes J. M. L. P. Caldeira, V. N. G. J. Soares, and P. D. Gaspar, "MobiSensA: Desenvolvimento de uma Aplicação Móvel para a Plataforma iSensA," in 13th Iberian Conference on Information Systems and Technologies (CISTI'2018), Cáceres, Spain, June 13-16, 2018.

[2] Posada, J., Toro, C., Barandiaran, I., Oyarzun, D., Stricker, D., et al, "Visual computing as a key enabling technology for Industry 4.0 and industrial Internet," IEEE Computer Graphics and Applications, 35(2), pp. 26-40, 2015.

[3] Roblek, V., Meško, M., Krapež, A, "A Complex View of Industry 4.0," SAGE Open 2016, 6(2), 2016.

Gruber, F.E, "Industry 4.0: A best practice project of the automotive industry," Digital Product and Process Development Systems, Springer Berlin Heidelberg, pp. 36-40, 2013.

[5] Vijaykumar, S., Saravanakumar, S.G., Balamurugan, M, "Unique sense: Smart computing prototype for Industry 4.0 revolution with IOT and bigdata implementation model," Indian Journal of Science and Technology, 8(35), 2015.

[6] Wan, J., Tang, S., Shu, Z., Li, D., Wang, S., et al, "Software-defined industrial Internet of Things in the context of Industry 4.0," IEEE Sensors Journal, 16 (22), pp. 7373-7380, 2016.

[7] Lu, Y., "Industry 4.0 - A survey on technologies, applications and open research issues," Journal of Industrial Information Integration, doi: 10.1016/j.jii.2017.04.005, 2017.

[8] Pragati, D., Kirtikumar, S. V., "IoT based Water Monitoring System: A Review," International Journal of Advance Engineering and Research Development, 4(6), 2017.

[9] Sukriti, Gupt, S., Indumathy, K, "IoT based Smart Irrigation and Tank Monitoring System," International Journal of Innovative Research in Computer and Communication Engineering, 4(9), 2016.

[10] Vaishnavi, D., Gaikwad, M, "Water Quality Monitoring System Based on IoT," Advances in Wireless and Mobile Communications, 10(5), pp. 1107-1116, 2017.

[11] Geetha, S., Gouthami, S, "Internet of things enabled real time water quality monitoring system," International Journal for @qua – Smart ICT for Water, 2(1), 2017.

[12] Kurde, A., Kulkarni, V, "IoT Based Smart Power Metering," International Journal of Scientific and Research Publications, 6(9), 2016.

[13] Pandit, S., Mandhre, S., Meghana, N. "Smart Energy Meter using Internet of Things," Vishwakarma Journal of Engineering Research, 1(2), 2017.

[14] Dharti, V., Borole, A., Singh, S., "Smart Agriculture Monitoring and Data Acqusition System," International Research Journal of Engineering and Technology, 3(3), 2016.

[15] Athani, S., Tejeshwar, C., Patil, M., Patil, P., Kulkarni, R., "Soil moisture monitoring using IoT enabled arduino sensors with neural networks for improving soil management for farmers and predict seasonal rainfall for planning future harvest in North Karnataka – India," Proceedings of the International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017), India, 10-11 February, pp. 43-48, 2017.
[16] Kumar, N., Vuayalakshmi, B., Prarthana, R., Shankar, A., "IoT based smart garbage alert system using Arduino UNO," Proceedings of the 2016 IEEE Region 10 Conference (TENCON), Singapore, 22-25 November, 2016.
[17] Revathy, S., Anandhi, A., Hemalatha, K., "IoT Based Smart Bin Monitoring Using Sensor and GSM for Smart Cities," International Journal of Research in Computer Science, 4(1), 2017.
[18] Patil, S., Vijayalashmi, M., "Solar Energy Monitoring System Using IoT," Indian Journal of Scientific Research, 15(2), pp. 149-155, 2017.
[19] Fioccola, G., Sommese, R., Tufano, I., Canonico, R., Ventre, G., "Polluino: An efficient cloud-based management of IoT devices for air quality monitoring," Proceedings of the IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), Bologna, Italy, 7-9 September, 2016.
[20] Patel, K., Patel, S., "IoT Based Data Logger for Monitoring and Controlling Equipment Working Status and Environmental Conditions," International Journal of Innovative Research in Computer and Communication Engineering, 4(4), 2016.
[21] Bouhai N. (Editor), S. Imad (Editor)., "Internet of Things: Evolutions and Innovations Volume 4," Wiley: NJ, USA, ISBN 9781119476573, 2017.
[22] Arduino. Available online: https://www.arduino.cc/ (accessed on 7 November 2017).
[23] Reneker, D., "Arduino vs. PLC for industrial control." In Control Design for Machine Builders, Jul 13, 2017. Available online: https://www.controldesign.com/articles/2017/arduino-vs-plc-for-industrial-control/ (accessed on 7 November 2017), 2017.
[24] ALBIGEC. Available online: http://www.albigec.pt/ (accessed on 21 September 2018).
[25] CM Pampilhosa da Serra. Available online: http://www.cm-pampilhosadaserra.pt/ (accessed on 21 September 2018).
[26] Serviços Municipalizados de Castelo Branco. Available online: http://www.sm-castelobranco.pt/ (accessed on 21 September 2018).
[27] Hermann, M.; Pentek, T.; Otto, B., "Design principles for Industry 4.0 scenarios," Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS), Institute of Electrical and Electronics Engineers - IEEE-, IEEE Computer Society, Kauai, Hawaii, 5-8 January, pp. 3928-3937, 2016.
[28] Thames, L., Schaefer, D., "Software-defined cloud manufacturing for Industry 4.0," Procedings CIRP, 52, pp. 12-17, 2016.
[29] Trappey, A.J.C., Trappey, C.V., Govindarajan, U.H., Chuang, A.C., Sun, J.J., et al., "A review of essential standards and patent landscapes for the Internet of Things: A key enabler for Industry 4.0," Adv. Eng. Informat, 2016.
[30] Queiroz, D.V., Alencar, M.S., Gomes, R. D., Fonseca, I.E., Benavente-Peces, C., "Survey and systematic mapping of industrial wireless sensor networks," Journal of Network and Computer Applications, 2017.
[31] iSensA. Available online: http://www.isensa.net/ (accessed on 21 September 2018).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 217-224 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Trajectory Tracking Control Optimization with Neural Network for Autonomous Vehicles

Samuel Oludare Bamgbose*, Xiangfang Li, Lijun Qian

Department of Electrical and Computer Engineering and CREDIT Center, Prairie View A & M University, Texas A & M University System, Prairie View, Texas, 77446, USA

ARTICLE INFO

Article history:
Received: 05 January, 2019
Accepted: 03 February, 2019
Online : 15 February, 2019

Keywords:
Nonlinear control
Optimal control
Neural network
Adaptive control
Autonomous systems
Mobile robot
2-DOF helicopter

ABSTRACT

For mission-critical and time-sensitive navigation of autonomous vehicles, controller design must exhibit excellent tracking performance with respect to the speed of convergence to reference command and steady-state accuracy. In this article, a novel design integration of the neural network with the traditional control system is proposed to adaptively obtain optimized controller parameters resulting in improved transient and steady-state performance of motion and position control of autonomous vehicles. Application of the proposed intelligent control scheme to mobile robot navigation was presented for an eight-shaped trajectory by optimizing a Lyapunov-based nonlinear controller. Furthermore, a Linear Quadratic Regulator-based controller was optimized based on the proposed strategy to control the pitch and yaw angles of a 2-Degree-of –Freedom helicopter. The simulation results showed that the proposed scheme outperforms the traditional controllers in terms of the speed of convergence to the desired trajectory and overall error minimization.

1. Introduction

There have been growing demands for autonomous vehicles with excellent maneuvering capabilities both for commercial and military applications [1, 2]. Significant amount of work is ongoing to realize the commercial operation of autonomous cars [3, 4], and Wheeled Mobile Robots (WMRs) are finding increasing use in industrial and service applications [5]. Autonomous Surface Vehicles (ASV) have been utilized to improve port safety and for ecological as well as meteorological purposes [6], whereas Autonomous Underwater Vehicles (AUVs) are useful tools for underwater search and inspection [7].

Safe navigation will require both excellent path planning and path tracking strategies. Path planning involves the determination of an appropriate trajectory for the vehicle whereas path tracking, which is the focus of this study, is the following of a desired trajectory. Several schemes have been reported in the literature for path planning, including in [8], where a framework was proposed to synthesize sequences of maneuvers that are tracked using nonlinear controllers, and other strategies are presented in [9, 10]. To execute the maneuvers generated by path planners, control strategies have been proposed as well.

In [11], Fuzzy Logic Controller (FLC) was proposed for robot tracking and Proportional-Integral-Derivative (PID) controller was utilized to control the speed of WMRs in [12]. The authors of [13] presented an observer-based approach, whereas backstepping control strategy was presented in [14, 15]. Sliding-mode control for WMR trajectory tracking with initial error was reported in [16], and modular-based method was proposed in [17]. The authors of [18] proposed a Model Predictive Control (MPC) approach and Kalman filter-based strategy was presented in [19]. Lyapunov-based scheme was reported in [20] and controller design by approximate linearization utilizing Taylor expansion was proposed in [21]. In order to enhance the perfomance of traditional control methods, Least Square Policy Iteration (LSPI) and Dynamic Heuristic Programming (DHP) algorithms were utilized in [22] for optimizing a Proportional-Derivative (PD) controller. Also, the authors of [23] utilized neural network (NN) to describe the inverse dynamics of a biped robot with respect to output errors for the control of level walking. However, application examples were limited to scenarios where the robot is initialized at the desired starting coordinate.

In order to execute maneuvers for unmanned Air Vehicles (UAVs), feedback linearization [24, 25] and sliding mode control [26] have been proposed, but they have failed for certain models due to the nonlinear dynamics of the vehicle [27]. For helicopter control, backstepping control strategy [28] and Linear Quadratic Regulator (LQR)-based control [29] have been proposed.

* Samuel Oludare Bamgbose Email: sobams77@yahoo.com

In this study, a novel learning-based adaptive scheme utilizing the neural network is developed for autonomous vehicle trajectory tracking. Whereas, plant models predict the vehicular motion for a given control command, accuracy is limited by modelling errors and approximations. Also, for certain models with complex nonlinear dynamics, it is difficult to obtain a suitable controller. Since machine learning models provide a more powerful tool to describe nonlinear dynamics of a plant given example data of plant operation, this article presents a parameterized control law designed to achieve trajectory tracking of autonomous vehicles adaptively. Rather than using a single controller, a family of controllers is obtained utilizing the NN model to estimate the time varying controller parameters. The scheme was applied to optimize a Lyapunov-based nonlinear controller parameters used to execute an eight-shaped maneuver for a mobile robot. Furthermore, LQR-based controller parameters for 2-Degree-Of-Freedom (DOF) helicopter position control were optimized using the proposed scheme. Simulation results show that the scheme outperforms the traditional control strategies in terms of faster convergence to the desired trajectory and more accurate steady-state performance, regardless of the initial shift in starting coordinates. Also, because the scheme is sample-based, it can compensate for modeling errors.

The rest of the paper is organized as follows; Section 2 presents the problem formulation. Section 3 provides an illustrative application of the proposed scheme. Simulation results are discussed in Section 4, and Section 5 summarizes the study.

2. Problem Formulation

A dynamic system can generally be defined in state-space as,

$$\dot{x}(t) = f(x(t), u(t)) \tag{1}$$

The system state is represented by $x(t)$, the control input is denoted by $u(t)$, whereas f denotes a mathematical function and $\dot{x(t)}$ is the state derivative with respect to time. Given a target signal $r(t)$, a traditional feedback control law can be computed based on the difference between the target signal and the actual system output. Such difference can be denoted by $e(t)$, resulting in a control input defined as,

$$u(t) = K(t)e\,(t) \tag{2}$$

where $K(t)$ denote the control gain, which can be constant or time-varying depending on the dynamic description of the system. However, by utilizing certain control strategies, not only the error is fed back for control but the states also. For two-dimensional and higher order systems; the states, control input and error signals are vectors, whereas the control gain could be a vector or matrix depending on the control strategy used.

In order to optimize the controller performance for changing system dynamics or operation regions, a neural network-based method that adaptively determines the control gain is proposed as follows. For a traditional closed-loop control system, sample test run of the system is performed and for each time step k, the state variables and the control gain are measured and arranged into 3-tuple $\{x(t), x(t+k), K(t)$. The state measurement before the application of control input is denoted as $x(t)$ whereas the next state caused by the control action is $x(t+k)$,. The control gain that caused the state transition is represented as $K(t)$.

Multiple samples of state transitions and control gains are measured for a sequence of operation, then the control gain or the control gain parameters for time-invariant or time-varying system respectively are varied for other sequences. Example of a sequence is the tracking of a particular course by a mobile robot or the change in angular orientation of a helicopter over a period of time. For training stability, the data set is normalized depending on the data range. And based on the accumulated data set as represented by the 3-tuple, a NN is trained using the state and next state as input and the control gain as the output. Hence,

$$K(t) = f(x(t), x(t+k)) \tag{3}$$

The NN model is made up of processing units called neurons with weighted interconnections among them [30], and the neurons with nonlinear activation functions are arranged into layers as shown in Figure 1.

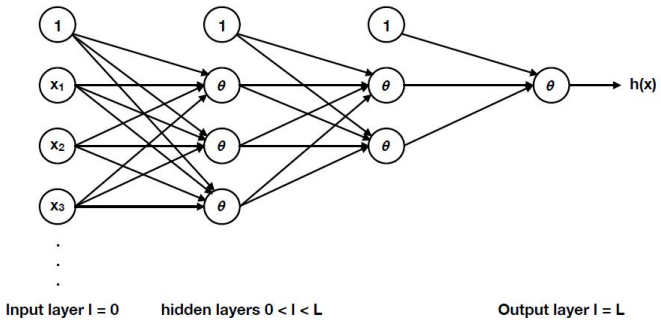


Figure 1: Feed forward neural network structure

The input layer is the vector of input variables, which are system states in the study. For a single hidden layer system, the hidden layer is a nonlinear combination of the input signals, and for multiple hidden layer system, subsequent layers are nonlinear combination of the previous layers. Knowledge is extracted from the output layer after model training. In supervised learning employed in this study, an error function based on the difference between the actual outputs called labels and the predicted outputs from the NN is minimized to adjust the interconnection weights among the neurons. After several iterations, the model is said to be trained. Whereas only a portion of the dataset is used for training, the remainder is used for testing the performance of the trained model. The mathematical description of the NN model as utilized in this study is,

$$K_n^{(l+1)} = \theta\left(s_n^{(l+1)}\right) \tag{4}$$

$$s_n^{(l+1)} = w^{(l+1)}K^l + b_n^{(l+1)} \tag{5}$$

where the output vector of a layer l is denoted by K^l. Hence, the input layer vector K^0 is equivalent to the state input $\{x\,(t), x(t+k)\}$. The activation function is denoted by θ and s^l represents the input vector to layer l. The interconnecting weights and biases from layer $l-1$ to unit n of layer l are represented by w_n^l and b_n^l respectively.

Following offline training as decribed above, the trained NN model is integrated as a control gain estimator in the traditional control system as shown in Figure 2.

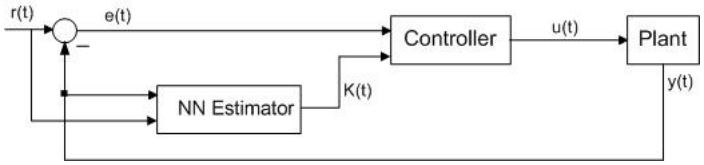


Figure 2: Neural network optimized adaptive control system

For online control gain estimation, the NN input due to the next state $x(t+k)$ is replaced with the reference signal $r(t)$. Hence,

$$K(t) = f(x(t), r(t)) \tag{6}$$

Then, (2) is transformed to,

$$u(t) = f\big(x(t), r(t)\big) * e(t) \tag{7}$$

3. Illustrative Examples

In order to prove the effectiveness of the proposed scheme, two application case studies are presented. One is the tracking of an eight-shaped trajectory by a mobile robot, and the other is the position tracking of a 2-DOF helicopter.

3.1. Mobile Robot Trajectory Tracking

The nonlinear kinetic model [31] of a non-holonomic wheeled mobile robot is described as,

$$\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\theta}\end{bmatrix} = \begin{bmatrix}cos\theta & 0\\ sin\theta & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}v\\ w\end{bmatrix} \tag{8}$$

The robot's states x, y, θ are the cartesian $x, y,$ and angular displacements respectively. Whereas $\dot{x}$, $\dot{y}$, and $\dot{\theta}$ are the state derivatives, v corresponds to the linear velocity and w represents the angular velocity around the vertical axis.

3.1.1. Traditional Control Design

Motion control of the mobile robot is achieved by computing appropriate linear and angular velocities to drive the robot in the desired trajectory. According to the procedure in [32], the state tracking error is generalized by defining it through a rotation matrix to obtain,

$$\begin{bmatrix}e_x\\ e_y\\ e_\theta\end{bmatrix} = \begin{bmatrix}cos\theta & sin\theta & 0\\ -sin\theta & cos\theta & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_d - x\\ y_d - y\\ \theta_d - \theta\end{bmatrix} \tag{9}$$

where x_d, y_d ,and θ_d are the target coordinates for the robot. Then the error dynamics in generalized coordinates is obtained by taking the derivative of (9) as,

$$\begin{bmatrix}\dot{e}_x\\ \dot{e}_y\\ \dot{e}_\theta\end{bmatrix} = \begin{bmatrix}0 & wd & 0\\ -wd & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}e_x\\ e_y\\ e_\theta\end{bmatrix} + \begin{bmatrix}vdcose_\theta - v\\ vdsine_\theta\\ wd - w\end{bmatrix} \tag{10}$$

Making the following substitustions,

$$u_1 = vdcose_\theta - v \tag{11}$$

$$u_2 = wd - w \tag{12}$$

The error dynamics is transformed to the subsequent form,

$$\begin{bmatrix}\dot{e}_x\\ \dot{e}_y\\ \dot{e}_\theta\end{bmatrix} = \begin{bmatrix}0 & w_d & 0\\ -w_d & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}e_x\\ e_y\\ e_\theta\end{bmatrix} + \begin{bmatrix}0\\ sine_\theta\\ 0\end{bmatrix}v_d + \begin{bmatrix}1 & 0\\ 0 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix} \tag{13}$$

To ensure global stability for the nonlinear dynamics, the following Lyapunov function is selected,

$$V = \frac{K_2}{2}\left(e_x^2 + e_y^2\right) + \frac{e_\theta^2}{2} \tag{14}$$

Taking the derivative of (14) gives

$$\dot{V} = K_2\left(e_x\dot{e}_x + e_y\dot{e}_y\right) + e_\theta\dot{e}_\theta \tag{15}$$

By substituting for $\dot{e}_x$, $\dot{e}_y$ and $\dot{e}_\theta$ from (13),

$$\dot{V} = K_2\left(e_x u_1 + e_y v_d sine_\theta\right) + e_\theta u_2 \tag{16}$$

Hence, the Lyapunov-based control law such that $\dot{V} < 0$ is guaranteed is obtained as,

$$u_1 = -K_1 e_x \tag{17}$$

$$u_2 = -K_2 v_d \frac{sine_\theta}{e_\theta} e_y - K_3 e_\theta \tag{18}$$

The controller gains are defined as,

$$K_1(t) = K_3(t) = 2\zeta\sqrt{w_d^2(t) + cv_d^2(t)} \tag{19}$$

where $K_2 = c > 0$ is a constant, and ζ is the damping ratio.

The control law defined by (17) and (18) are composed of both the feedback and feedforward signals. The feedforward commands are determined from (8) as,

$$v_{d(t)} = \pm\sqrt{\dot{x}_d^2(t) + \dot{y}_d^2(t)} \tag{20}$$

$$w_d(t) = \frac{\ddot{y}_d(t)\dot{x}_d(t) - \dot{x}_d(t)\dot{y}_d(t)}{\dot{x}_d^2(t) + \dot{y}_d^2(t)} \tag{21}$$

$$\theta_d(t) = atan2(\dot{y}_d(t), \dot{x}_d(t)) + \pi i \tag{22}$$

where i is 0 or 1 for forward or backward motion respectively. The feedback commands are adaptively computed based on NN estimation as described next.

3.1.2. Controller Parameter Estimation using the Neural Network

In order to optimize the controller for fast convergence to the desired trajectory and improved steady-state performance, the NN is utilized to adaptively estimate the feedback control parameters. Sample runs of the traditional closed-loop system are performed for 15 sequences of 1258 state transitions to make a total of 18870 data samples. The input features are the current state$[x(t)\ \ y(t)\ \ \theta(t)]$, the nominal velocities $[v_{d(t)}\ \ w_{d(t)}]$, and

the next state$[x(t+k)\ y(t+k)\ \theta(t+k)]$. The output labels are the control gains $[K_1(t)\ \ K_2(t)\ \ K_3(t)]$ responsible for the state transition. Data was normalized and the training was implemented using the MATLAB NN tool box using 70% of the dataset. The remaining dataset was divided into 15% validation and 15% testing. The Levenberg-Marquerdt algorithm [33, 34, 35] was used for training using two hidden neurons. The performance of the NN estimator was evaluated using the Mean Square Error (MSE) plot shown in Figure 3. The figure shows good generalization performance as the test error is approximately equal to that of the training.

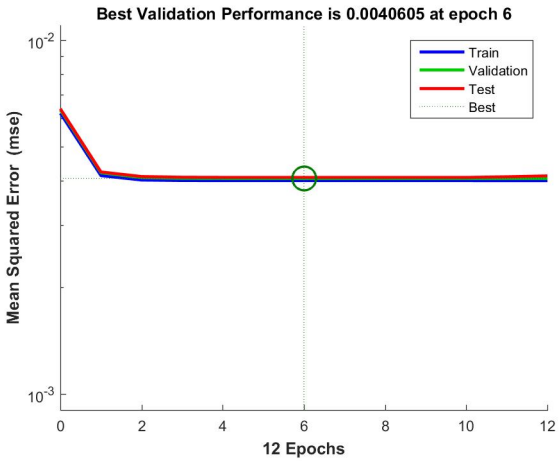


Figure 3: MSE of the NN estimator for robot control

The trained NN estimator is integrated in the second loop of Figure 2, where a normalizer and denormalizer are embedded for better control performance. Simulation results of the adaptive closed loop control system is presented in Section 4.

3.2. 2-DOF Helicopter Position Control

A Two-Input-Two-Output (TITO) Quanser 2D Helicopter model [35] is considered. The linearized model is obtained in [36] as follows.

$$\begin{bmatrix}\dot{\theta}\\ \dot{\varphi}\\ \ddot{\theta}\\ \ddot{\varphi}\end{bmatrix}=\begin{bmatrix}0 & 0 & 1.0000 & 0\\ 0 & 0 & 0 & 1.0000\\ 0 & 0 & -0.0315 & 0\\ 0 & 0 & 0 & -0.0673\end{bmatrix}\begin{bmatrix}\theta\\ \varphi\\ \dot{\theta}\\ \dot{\varphi}\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & 0\\ 0.5293 & 0.0404\\ 0.0372 & 0.1333\end{bmatrix}\begin{bmatrix}V_{mp}\\ V_{my}\end{bmatrix} \quad (23)$$

where θ and φ are the pitch and yaw angles respectively.The first and second derivatives of the pitch and yaw angles are $\dot{\theta},\ \ddot{\theta}\ and\ \dot{\varphi},\ \ddot{\varphi}$ respetively. The pitch and yaw propeller voltages V_{mp} and V_{my} are applied to the corresponding motors driving the propellers.

3.2.1. Traditional Control Design

The pitch and yaw angles of the 2-DOF helicopter were regulated using the optimal LQR-based control design presented in [29]. And the control law was obtained as,

$$u = K_x x(t) - K_e e(t) \quad (24)$$

The realized state feedback gains are,

$$K_x = \begin{bmatrix}547.8 & -2.8 & 45.7 & -4.8\\ 2.8 & 547.9 & -4.1 & 91\end{bmatrix} \quad (25)$$

and the error feedback gains were obtained as,

$$K_e = \begin{bmatrix}-1.0000 & 0.0051\\ -0.0051 & -1.000\end{bmatrix} \quad (26)$$

The system states are represented by $x(t)$ whereas $e(t)$ is the error feedback signal. The adaptive computation of the error feedback gains using the NN is subsequently described.

3.2.2. Controller Parameter Estimation using the Neural Network

In order to achieve better accuracy and faster tracking convergence, the NN is employed to optimize the error feedback control gains. In this case, six sequences of 700 state transitions totaling 4195 data samples were sufficient, since the plant has been linearized. The input features are the current angular orientation $[\theta(t)\ \varphi(t)]$ along with the next angular orientation $[\theta(t+k)\ \varphi(t+k)]$ of the 2-DOF helicopter. The output labels are the error feedback control gains $[K_{e11}(t)\ \ K_{e12}(t)\ \ K_{e21}(t)\ \ K_{e22}(t)]$ responsible for the angular orientation change. Similar procedures as proposed in Subsession 3.1.2 of this paper are followed, and the MSE plot is as shown in Figure 4. The figure shows good generalization of the estimator similar to observation in the previous section since the test error approximates the training error. The trained NN estimator is integrated in the second loop of Figure 2 to obtain optimized performance as presented next in Section 4.

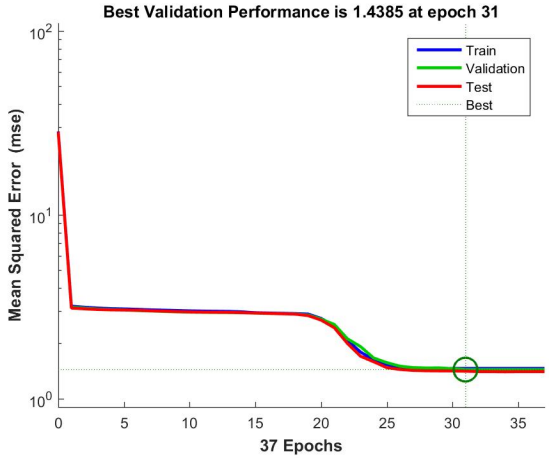


Figure 4: MSE of the NN estimator for helicopter control

4. Simulation Results

Closed-loop simulation was conducted in Matlab Simulink for both the traditional and the NN-optimized control system. The results for different initial state shifts were compared to prove the effectiveness of the proposed scheme. The desired eight-shaped trajectory for the mobile robot was defined as,

$$xd(t) = sin\frac{t}{10} \quad (25)$$

$$yd(t) = sin\frac{t}{20} \quad (26)$$

$$\theta d(t) = atan2\ (y\dot{d}(t), \dot{x}d(t)) \quad (27)$$

The feed-forward driving and steering velocities as described by (20) and (21) are shown in Figures 5 and 6 respectively.

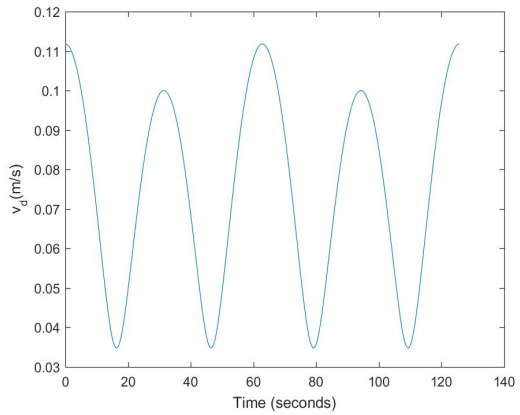


Figure 5: Robot feed-forward driving velocity input

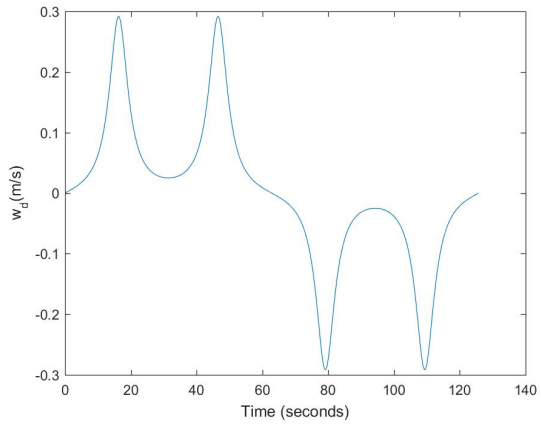


Figure 6: Robot feed-forward steering velocity input

Figures 7 – 9 is the trajectory tracking performance of the nonlinear Lyapunov-based controller for the mobile robot with initial state error $[0.1, 0.3, 0]$, where the linear displacement unit is meters and radians is the unit of the angular displacement. Figure 7 is the result of an underdamped control system due to low damping ratio, whereas Figure 8 is the result of an over-damped control system due to high damping ratio. The result of a critically damped control system with unity damping ratio is shown in Figure 9, whereas Figure 10 is the corresponding trajectory tracking output performance of the NN-optimized control system. It can be observed that the NN-optimized controller outperformed the different design modes of Lyapunov-based controller in terms of faster convergence to the desired trajectory and steady-state error minimization. The steady-state is the region starting at the point where the robot settles around the reference trajectory. Although, either fairly good convergence or steady-state performance can be obtained by varying the control parameters of the Lyapunov-based controller, but the improvement is mutually exclusive. However, the NN-optimized controller achieved both better transient and steady state performances concurrently by adaptively obtaining varying controller parameters.

Figures 11 - 14 are the control performance outputs for the mobile robot trajectory tracking with initial state error $[0.3, -0.5, {}^{\pi}/_{6}]$. Also in this case, linear and angular displacements are measured in meters and radians respectively. It can be observed, as the case is in the previous example, that the NN-optimized controller of Figure 14 showed better comparative control performance than the different variations of traditional Lyapunov-based controllers of Figures 11 – 13.

To further show the performance advantage of the NN-optimized controller, the average linear and angular tracking errors over the entire trajectory are presented in Table 1 as it applies to the different initial state errors. It can be seen that both the resultant linear and angular tracking errors are minimized using the NN-optimized controller.

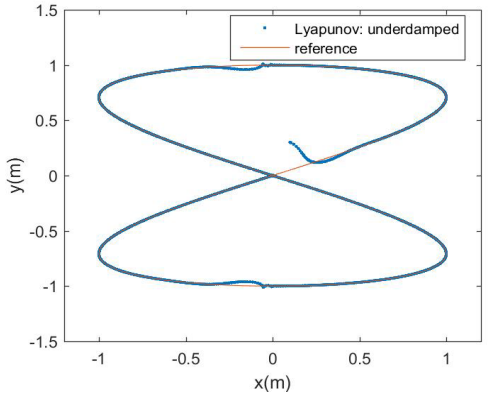


Figure 7: Underdamped robot Lyapunov-based control

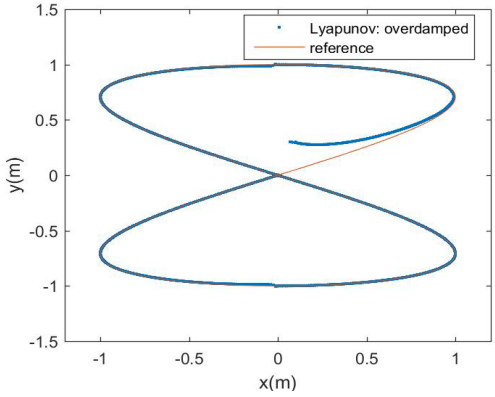


Figure 8: Overdamped robot Lyapunov-based control

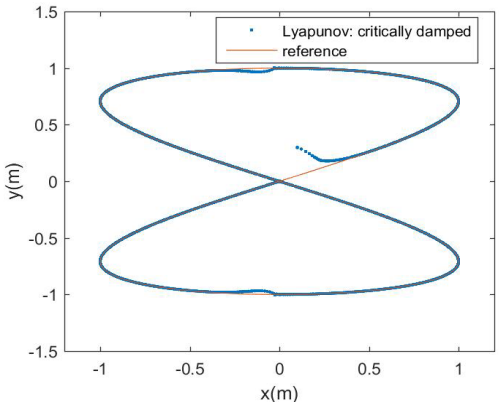


Figure 9: Critically damped robot Lyapunov-based control

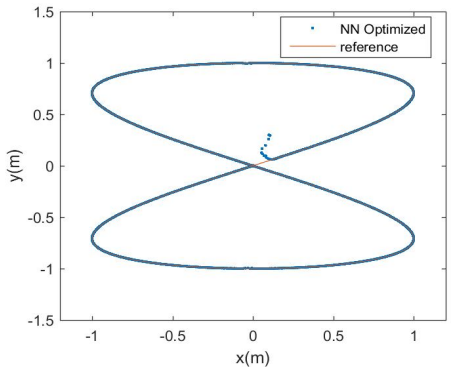


Figure 10: Robot NN-optimized control

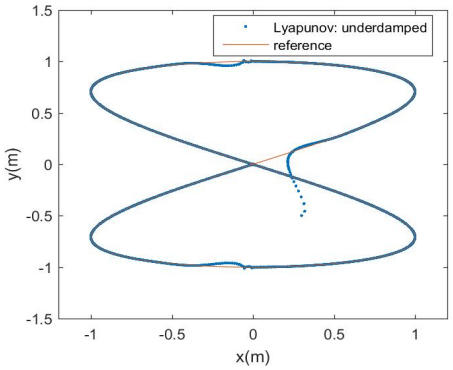


Figure 11: Underdamped robot Lyapunov-based control

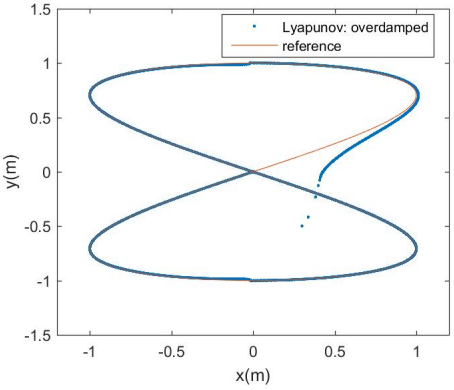


Figure 12: Overdamped robot Lyapunov-based control

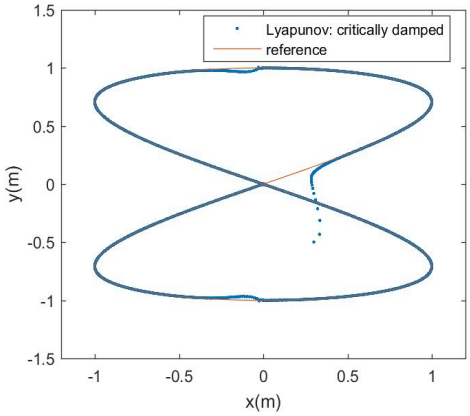


Figure 13: Critically damped robot Lyapunov-based control

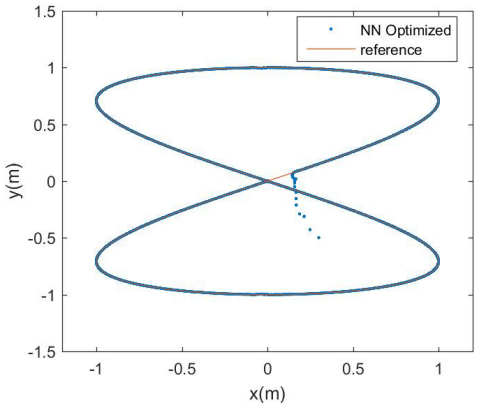


Figure 14: Robot NN-optimized control

Table 1: Robot mean tracking error based on control method

Control Method	Initial state error [m,m,rad]	Mean x error (m)	Mean y error (m)	Mean θ error (m)
Lyapunov_ underdamped	[0.1,0.3,0]	1.7×10^{-3}	-2.3×10^{-3}	2.4×10^{-2}
Lyapunov_ overdamped	[0.1,0.3,0]	-1.8×10^{-4}	-9.2×10^{-3}	2.3×10^{-2}
Lyapunov_ critically damped	[0.1,0.3,0]	2.8×10^{-4}	-3.8×10^{-3}	2.6×10^{-2}
Neural network optimized	[0.1,0.3,0]	3.5×10^{-4}	-2.5×10^{-4}	2.2×10^{-2}
Lyapunov_ underdamped	[0.3,-0.5,π/6]	3.5×10^{-3}	2.1×10^{-3}	-2.1×10^{-2}
Lyapunov_ overdamped	[0.3,-0.5,π/6]	4.4×10^{-4}	1.33×10^{-2}	-3.5×10^{-2}
Lyapunov_ critically damped	[0.3,-0.5,π/6]	1.4×10^{-3}	4.1×10^{-3}	-2.3×10^{-2}
Neural network optimized	[0.3,-0.5,π/6]	-5.1×10^{-4}	-1.3×10^{-4}	-3.3×10^{-2}

The desired positioning of the 2-DOF helicopter pitch and yaw angular orientations over time are defined as,

$$\theta_d(t) = \begin{cases} 0.8\tau^5 - 8\tau^3 + 10\tau + 6, \tau = 0.05t, t \leq 47 \\ -16.9, \ t > 47 \end{cases} \quad (28)$$

$$\varphi_d = 5 \quad (29)$$

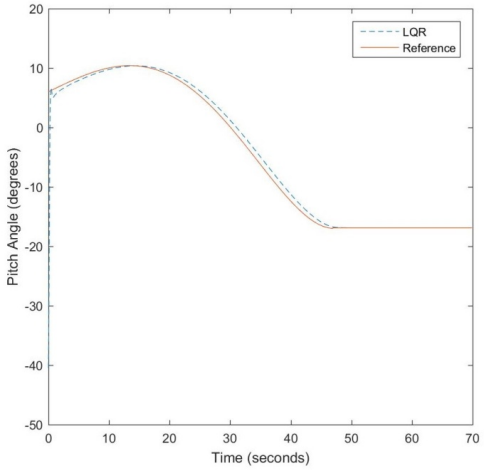


Figure 15: LQR-based pitch angle control

The LQR-based controller performance is shown in Figures 15 and 17 with initial pitch and yaw angular orientations of -40.5 and 0 degrees. The Q and R parameters were selected as $diag([3e5\ 3e5\ 1\ 1\ 1\ 1])$ and $diag([1\ 1])$ where Q is associated with both the state and error feedback signals. The performances of the NN-optimized controller are shown in Figures 16 and 18 for pitch and yaw angle control respectively. Since the system is linearized, modest performance improvement is observed with respect to response speed and steady-state error minimization. Table 2 shows the comparative tracking errors.

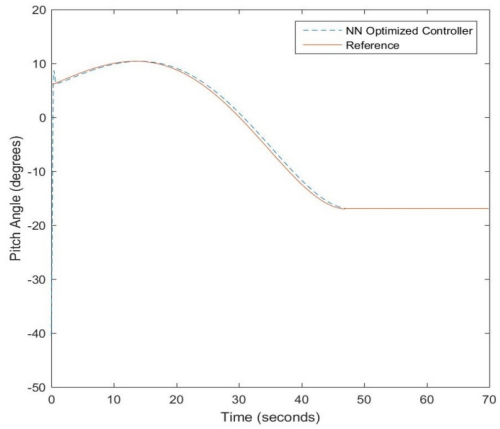


Figure 16: NN-optimized pitch angle control

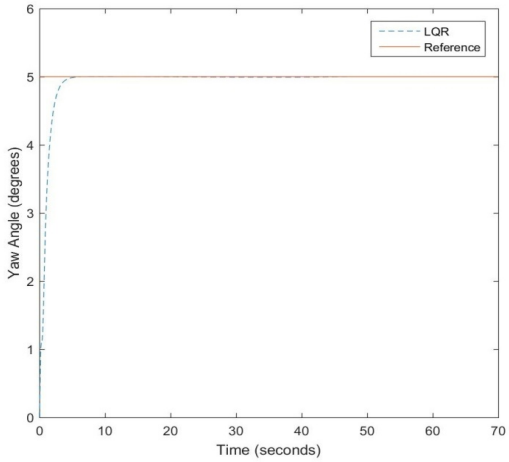


Figure 17: LQR-based yaw angle control

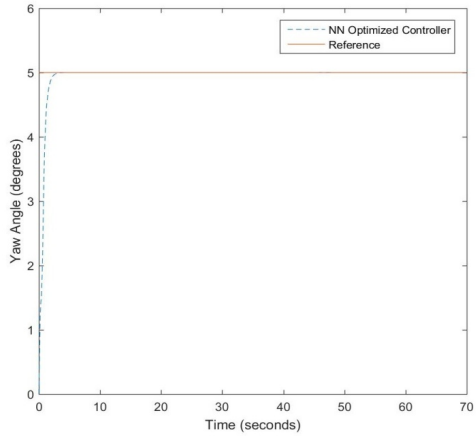


Figure 18: NN-optimized yaw angle control

Table 2: Helicopter mean tracking error based on control method

Control Method	Initial state error (degrees)	Mean θ error (degrees)	Mean φ error (degrees)
LQR	[-40.5,0]	-0.18	0.08
Neural network optimized	[-40.5,0]	-0.06	0.04

5. Conclusion

A neural network optimized control system design for autonomous vehicle navigation has been proposed in this study. The design consists of an inner error loop integrated with an outer loop for estimating the controller parameters utilizing a neural network trained on samples from test navigation. Comparative studies of the proposed scheme with traditional methods were presented. In the first case, the neural network optimized control system was shown to outperform a Lyapunov-based controller in terms of faster convergence to the desired trajectory and better steady-state performance for an eight-shaped maneuver. A second illustrative simulation example was conducted for the control of pitch and yaw angles of a 2-Degree-Of-Freedom helicopter model, where improved transient and steady-state performances were observed for NN-optimized controller over a Linear Quadratic Regulator-based controller.

Because the neural network structure allows complex and nonlinear mapping of variables, the trained estimator is able to learn the behavior of both linear and nonlinear systems. Furthermore, since the proposed scheme is sample-based, it can compensate for plant modelling errors that degrades the performance of traditional controllers in real-life applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research work presented in this article is supported in part by the US National Science Foundation award 1464387, 1736196 and by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the US National Science Foundation or the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

References

[1] S. Thrun, M. Montemerlo, H. DahlKamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffman et al, "Stanley: The robot that won the DARPA grand challenge". J. Field Robot, **23**(9), 661-692, 2006

[2] D. Weatherington, Unmanned Aerial Vehicle Roadmap Report, The Information Warfare Site, 2003. http://www.iwar.org.uk/news-archive/2003/03-18-5.htm

[3] T. Luettel, M. Himmelsbach, H. J. Wuensche, "Autonomous ground vehicles-concepts and a path to the future" in Proc. of the IEE, **100** (Special Cantennial Issue), 1831-1839, 2012. https://doi.org/10.1109/JPROC.2012.2189803

[4] T. Drage, J. Kalinowski, T. Braunl, "Integration of drive-by-wire with navigation control for a driverless electric race car" IEEE Intelligent Transportation Systems Magazine, **6**(4), 23-33, 2014. https://doi.org/10.1109/MITS.2014.2327160
[5] R. D. Schratt, G. Schmierer, Serviceroboter, Springer, Berlin, 1998
[6] N. Miskovic, D. Nad, L. Rendulic, "Tracking divers: An autonomous marine surface vehicle to increase diverse safety" IEEE Robotics Automation Magazine, **22**(3), 72-84, 2015. https://doi.org/10.1109/MRA.2015.2448851
[7] F. Plumet, C. Petres, M.A. Romero-Ramirez, B. Gas, S.H Long, "Toward and autonomous sailing boat" IEEE Journal of Oceanic Engineering, **40**(2), 397-407, 2015. https://doi.org/10.1109/JOE.2014.2321714
[8] E. Frazzoli, M.A. Dahleh, E. Feron, "Maneuver based motion planning for nonlinear systems with symmetries" IEEE Trans. Robot. Automat., **21**, 1077-1091, 2005. https://doi.org/10.1109/TRO.2005.852260
[9] C. Dever, B. Mettler, E. Feron, J. Popovic, M. McConley, "Nonlinear trajectory generation for autonomous vehicles via parameterized maneuver classes" J. Guid. Control Dyn., **29**(2), 289-302, 2006 https://doi.org/10.2514.1.13400
[10] N. K. Ure, G. Inalhan, "Autonomous Control of unmanned combat air vehicles: design of a multimodel control and flight planning framework for agile maneuvering" IEEE Control Systems Magazine, **32**(5), 74-95, 2012. https://doi.org/10.1109/MCS.2012.2205532
[11] S. Qiu, Z. Li, W. He, L. Zhang, C. Yang, C. Y. Su, "Brain-machine interface and visual compressive sensing-based teleoperation control of an exoskeleton robot" IEEE Trans. On Fuzzy Systems, **25**(1), 58-69, 2017. https://doi.org/10.1109/TFUZZ.2016.2566676
[12] C. S. Shijin, K. Udayakamar, "Speed control of wheeled mobile robots using pid with dynamics and kinetic modelling" in 4th International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 1-7, 2017. https://doi.org/10.1109/ICIIECS.2017.8275962
[13] M. Chen, "Disturbance attenuation tracking control for wheeled mobile robots with skidding and slipping" IEEE Trans. On Industrial Electronics, **64**(4), 3359-3368, 2017. https://doi.org/10.1109/TLE.2016.2613839
[14] W. M. E. Mahgoub, I. M. H. Sanhoury, "Back stepping tracking controller for wheeled mobile robot" in Proc. of International Conference on Communications, Control, Computing and Electronics Engineering (ICCCCEE), 1-5, 2017. https://doi.org/10.1109/ICCCCEE.2017.7867663
[15] O. Chwa, "Slidding-mode tracking control of nonholonomic wheeled mobile robots in planar corrdinates" IEEE Trans. Control Syst. Technol., **12**, 637-666, 2004. https://doi.org/10.1109/TCST.2004.824953
[16] Z. Jiang, H. Nijmeijer, "Tracking control of mobile robots: a case study in backstepping" Automatica, **33**, 1393-1399, 1997 https://doi.org/10.1016/50005-1098(97)00055-1
[17] M. M. Michalek, "Cascade-like modular tracking controller for non-standard N-Trailers" IEEE Trans. On Control Systems Technology **25**(2), 619-627, 2017 . https://doi.org/10.1109/TCST.2016.2557232
[18] H. Yang, M. Guo, Y. Xia, L. Cheng, "Trajectory tracking for wheeled mobile robots via model predictive control with softening constraints". IET Control Theory and Applications, **12**(2), 206-214, 2018. https://doi.org/10.1049/iet.cta.2017.0395
[19] M. Cui, W. Liu, H. Liu, X. Lu, "Unscented Kalman filter-based adaptive tracking control for wheeled mobile robots in the presence of wheel slipping" in Proc. 12th World Congress on Intelligent Control and Automation (WCICA), 3335-3340. https://doi.org/10.1109/WCICA.2016.7578636
[20] Y. Wang, Z. Miao, H. Zhong, Q. Pan, "Simultaenous stabilization and tracking of nonholonomic mobile robots: a lyapunov-based approach" IEEE Trans. On Control Systems Techonology, **23**(4), 1440-1450, 2015 https://doi.org/10.1109/TCST.2014.2375812
[21] F. Kuhne, W. F. Lages, J. M. G. da Silva, L. Jono "Model predictive control of a mobile robot using linearization" in Proc. Mechatronics and Robotics, 2004.
[22] H. Yang, Q. Guo, X. Xu, C. Lian, "Self-learning pd algorithms based on approximate dynamic programming for robot motion planning" in 2014 International Joint Conference on Neural Networks (IJCNN), 3663-3670, 2014. https://doi.org/10.1109/IJCNN.2014.6889711
[23] J. K. Rai, V. P. Singh, R. P. Tawari, D. Chandra, "Artificial Neural Network controllers for biped robot" 2nd International Conference on Power, Control and Embedded Systems, 625-630, 2012. https://doi.org/10.1109/ICPCES.2012.6508093
[24] A. Isidori, Nonlinear Control Systems, M. Thoma, Ed., Springer NewYork, 1995.
[25] S. Sastry, Nonlinear Systems, J. E. Marsden, L. Sirovich, S. Wiggins, Eds., Springer New York, 1999.
[26] V. I. Utkin, Sliding Modes in Control and Optimization, Springer Berlin, 1992
[27] J. J. E Slobine, W. Li, Applied Nonlinear Control, Prentice Hall, Englewood Cliffs, NJ, 1991
[28] E. Frazzoli, M. Dahleh, E. Feron, "Trajectory tracking control design for autonomous helicopters using backstepping algorithm" in Proc. American Control Cont., **6**, 4102-4107, 2000. https://doi.org/10.1109/ACC.2000.876993
[29] X. Meng, Y. Zhang, J. Zhang, "Multiple input/ouput delays compensation for networked mimo systems" in 25th Chinese Control and Decision Conference (CCDC), 2545-2550, 2013. https://doi.org/10.1109/CCDC.2013.6561369
[30] Y. S. Abu-Mostafa, M. Magdon-Ismail, H. Lin, Learning from Data, AMC book.com, USA, 2012.
[31] S. Blazic, "Two approaches for nonlinear control of wheeled mobile robots" in 13th IEEE International Conference on Control Automation (ICCA), 946-951, 2017. https://doi.org/10.1109/ICCA.8003188
[32] A. D. Luca, G. Oriolo, M. Venditteli, Control of Wheeled Mobile Robots: An Experimental Overview In: S. Nicosia, B. Siciliano, A. Bicchi, P. Valigi (eds) Ramsete. Lecture Notes in Control and Information Sciences, **270**, Springer, Berlin, Heidelberg, 2001.
[33] K. Levenberg, "A method for the solution of certain problems in least square" Quart. Appl. Math., **2**, 164-168, 1944. https://doi.org/10.1109 /qam/10666
[34] D. Marquerdt, "An algorithm for least-square estimation of linear parameters" SIAM J. Appl. Math., **11**(2), 431-441, 1963. https://dx.doi.org/10.1137/0111030
[35] A. Ranganathan, "The Levenberg-Marquardt Algorithm". Honda Research Institute USA, 2004. http://ananth.in/Notes-files/Imtut.pdf
[36] Quanser 2 DOF Helicopter User and Control Manual. Quanser Inc., Canada, 2006.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 225-230 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

PAPR and BER Performances of OFDM System with Novel Tone Reservation Technique Over Frequency Non-Selective Fading Channel

Moftah Ali*, Raveendra K. Rao, Vijay Parsa

Innovation Centre for Information Engineering (ICIE), Faculty of Engineering, Department of Electrical and Computer Engineering, University of Western Ontario, London, Ontario, N6A 5B9, Canada

A R T I C L E I N F O

Article history:
Received: 21 December, 2018
Accepted: 08 February, 2019
Online : 20 February, 2019

Keywords:
Orthogonal Frequency Division Multiplexing
Tone Reservation
Pseudo-random Sequence

A B S T R A C T

An Orthogonal Frequency Division Multiplexing (OFDM) system with Quadrature Phase Shift Keying (QPSK) mapper is considered. A novel low-complexity Tone Reservation (TR) technique is proposed for reduction of Peak-to-Average Power Ratio (PAPR) of the system. The technique is easy-to-implement and minimizes the search space of phases of reserved tones in the system. The ability of PAPR reduction of this proposed TR technique is assessed and compared with conventional TR technique that uses Pseudo Noise (PN) sequences to determine phases of reserved tones. The simulation results illustrate that the proposed TR technique is nearly the same as that of the conventional TR technique in terms of PAPR reduction capability but with reduced complexity. The Bit Error Rate (BER) performance of QPSK-OFDM system with the novel TR scheme over frequency non-selective Rayleigh fading channel is also determined and illustrated.

1. Introduction

Multicarrier modulation such as OFDM is extensively adopted in numerous communication systems and standards. For example, Digital Audio Broadcasting (DAB) [1], IEEE set of standards that includes wireless local area network 802.11a [2], wireless broadband standard 802.16 [3], and mobile broadband wireless access 802.20 [4]. In addition, OFDM is the core technique for 3GPP [5] and Large Term Evolution (LTE) [6] etc. OFDM systems are attractive due to: i) their immunity to multipath fading and impulse noise; ii) elimination of complex equalizer; and iii) efficient implementation using IFFT/FFT for modulation/demodulation. OFDM system has one significant drawback that is the transmitted signals exhibit high values of PAPR. As a result, a High Power Amplifier (HPA) with sufficient linear amplification range is required in the system; otherwise, the transmitted signal would be nonlinearly distorted resulting in loss of subcarrier orthogonality, and hence BER degradation of the system. Thus, it is important to consider an appropriate PAPR reduction technique in an OFDM for it to be cost effective as well as energy efficient. In the literature, several methods have been proposed and investigated, which can mitigate PAPR problem in an OFDM system, such as Clipping and Filtering (CF) [7], weighting [8], Selective Mapping (SLM) [9], [10], Active Constellation Extension (ACE) [11], Partial Transmit Sequence [12], [13], Tone Reservation (TR) [14], Tone Injection [14], coding [15] etc. Each technique used in an OFDM system offers its own computational complexity, PAPR reduction capability, amount to side information required to be exchanged between transmitter and receiver, and BER performance of the system [16], [17], [18].

The focus in this paper is on TR technique for reduction of PAPR in an OFDM system that requires: i) fewer computations compared to TR techniques available in the literature and ii) no side information to be exchanged between transmitter and receiver. This method was briefly presented in [19]. In this paper, this work is extended and examined in more detail and is compared with the well-known SLM technique for PAPR reduction [18]. It is noted that SLM is a distortion-less technique and requires side information (SI) to be conveyed to receiver. In addition, the QPSK-OFDM system with TR method introduced in the paper is examined for BER performance over Rayleigh fading channel.

In the paper, a novel TR technique with low-complexity is presented for reduction of PAPR in an OFDM system; the proposed scheme is easy-to-use and minimizes the search space of phase sequences. The technique achieves approximately same performance as that of TR technique using conventional PN

*Corresponding Author: Moftah Ali, Email: mali254@uwo.ca

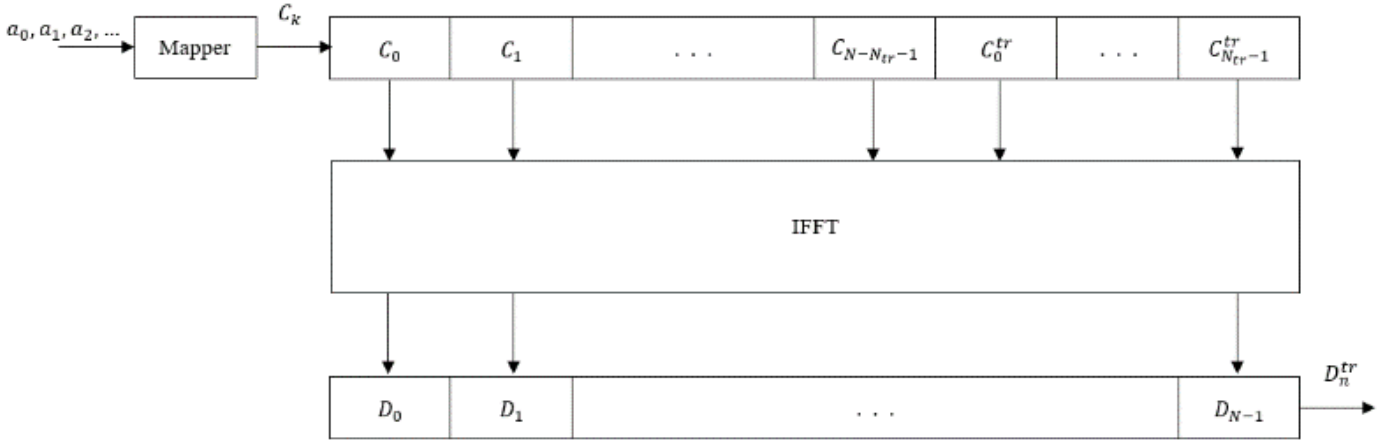


Figure 1: Block diagram for generation of OFDM signal with TR technique.

sequence. The simulation results are provided for a 128-subcarrier QPSK-OFDM system, as a function of numbers of tones reserved The probability of bit error performance of the system is also presented.

In section 2, OFDM signal and conventional TR technique are explained. Section 3 describes the TR method introduced in the paper. In section 4, OFDM receiver is discussed and in section 5, BER of the system in frequency-non-selective Rayleigh fading channel is presented. In section 6, numerical results are presented and discussed. The paper concluded in section 7.

2. OFDM Signal Model and TR Technique

2.1. OFDM Signal and its PAPR

An OFDM signal is a summation of N (number of subcarriers or tones) independently modulated complex sinusoidal signals, and is given by:

$$D(t) = \frac{1}{N}\sum_{k=0}^{N-1} C_k\, e^{j2\pi f_k t}, \qquad 0 \le t \le T_s \tag{1}$$

where $j = \sqrt{-1}$, T_s is the symbol duration and, $C_k, k = 0, 1, \ldots, N-1$, are the complex data. When QPSK mapper is considered, $C_k \in \{\pm 1 \pm j\}$ and the subcarrier frequencies are $f_k = k/T_s$, $k = 0, 1, \ldots, N-1$. When signal in (1) is sampled at Nyquist rate, discrete OFDM signal obtained and is given by:

$$D_n = D(n2T_b) = \tfrac{1}{N}\textstyle\sum_{k=0}^{N-1} C_k\, e^{\frac{j\pi nk}{N}},\ n = 0, 1, \ldots N-1 \tag{2}$$

The PAPR of signal in (1) is:

$$PAPR = \frac{\underset{0 \le t \le T_s}{max}\ |D(t)|^2}{\frac{1}{T_s}\int_0^{T_s} |D(t)|^2\, dt}, \tag{3}$$

Equivalently, PAPR in (3) is given by:

$$PAPR = \frac{max\{|D_n|^2, n = 0, 1, \ldots, LN-1\}}{\frac{1}{NL}\sum_{n=0}^{NL-1} |D_n|^2}, \tag{4}$$

where L (≥ 4) is the oversampling factor.

2.2. TR Technique

In TR method, N_{tr} of N tones, are reserved and used for PAPR reduction [20]. Figure 1 illustrates a block diagram for generation of OFDM signal with TR technique. The ratio $R = N_{tr}/N$ is typically kept small. Thus, the OFDM signal with tone reservation technique can be expressed as:

$$D_n^{TR} = D_n + D_n^{tr} \tag{5}$$

$$D_n^{TR} = \frac{1}{N}\sum_{k=0}^{N-1} (C_k + C_k^{tr})\, e^{j2\pi kn/N}, \quad n = 0, 1, \ldots, N-1 \tag{6}$$

where $C_k = [C_0, C_1, \ldots, C_{N-1}]$ and $C^{tr} = [C_0^{tr}, C_1^{tr}, \ldots, C_{N-1}^{tr}]$ represent frequency domain vectors associated with data and reserved tones. The reserved tones' locations are denoted by the index $\mathcal{R}^{tr} = \{i_0, i_1, \ldots, i_{N_{tr}-1}\}$ where $0 \le i_0 \le i_1 \le \cdots \le i_{N_{tr}-1} \le N-1$. Let the complement of $\mathcal{R}^{tr}$ be the index set $\mathcal{R}$ in $\mathcal{N} = \{0, 1, \ldots, N-1\}$. Therefore, $C_k \equiv 0\ \forall\ k \in \mathcal{R}^{tr}$ and $C_k^{tr} \equiv 0\ \forall\ k \in \mathcal{R}$. This can be expressed as:

$$C_k + C_k^{tr} = \begin{cases} C_k, & k \in \mathcal{R} \\ C_k^{tr}, & k \in \mathcal{R}^{tr} \end{cases} \tag{7}$$

The PAPR is redefined as:

$$PAPR^{tr} = \frac{max|D_n + D_n^{TR}|^2}{E\{|D_n|^2\}} \tag{8}$$

The objective in tone reservation technique is to obtain $\{C_k^{tr}, k = 0, 1, \ldots, N-1\}$ such that:

$$C^{tr(opt)} = \underset{(C^{tr} \in a_{mk})}{min}\ max\{|D_n^{TR}|^2, n = 0, 1, \ldots, N-1\} \tag{9}$$

3. Proposed Phase Sequence Scheme

The PN sequence is generated based on concatenating matrices. Denoting the seed matrix by $\boldsymbol{A}_0$ that can be expressed as:

$$\boldsymbol{A}_0 = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

where $a_i \in \{1, -1, +j, -j\}$, $i = 1, 2, 3, 4$ and $\boldsymbol{A}_0$ must include all the elements in the given set. Thus, one possible realization of $\boldsymbol{A}_0$ is:

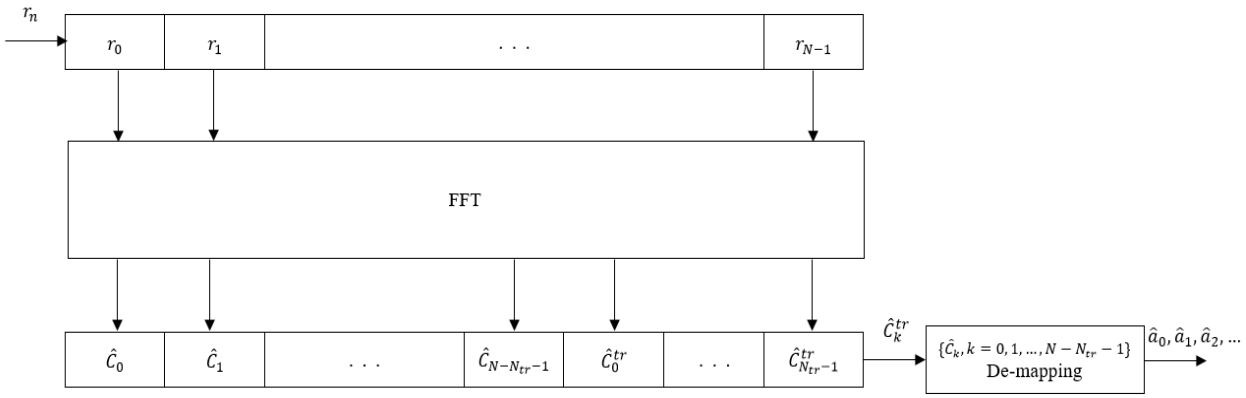


Figure 2: Block diagram for demodulation of OFDM signal with TR technique.

$$\boldsymbol{A}_0 = \begin{bmatrix} 1 & j \\ -1 & -j \end{bmatrix}$$

It was reported in [19] that not all available patterns of $\boldsymbol{A}_0$ will perform the same. Figure 3 demonstrates a sample of the possible patterns of the seed matrix $\boldsymbol{A}_0$.

a	b	c
$\begin{bmatrix} 1 & -j \\ -1 & j \end{bmatrix}$	$\begin{bmatrix} j & 1 \\ -1 & -j \end{bmatrix}$	$\begin{bmatrix} 1 & -j \\ j & -1 \end{bmatrix}$

Figure 3: Sample of the available patterns of $\boldsymbol{A}_0$.

In Figure 3, pattern (c) must be avoided for $\boldsymbol{A}_0$ while Pattern (b) performs poorer than pattern (a).

It is noted that phase sequences in the proposed TR technique can be generated using seed matrix $\boldsymbol{A}_0$, which is controlled by:

$$\boldsymbol{A}_m = \begin{bmatrix} \boldsymbol{A}_{m-1} & \boldsymbol{A}_{m-1} \\ \boldsymbol{A}_{m-1} & \boldsymbol{A}_{m-1}^T \end{bmatrix}, \qquad m = \log_2 N - 1 \quad (10)$$

where, $\boldsymbol{A}_{m-1}^T$ is conjugate transpose of $\boldsymbol{A}_{m-1}$ and $N = 2^n$, $n = 2, 3, 4, \ldots$

4. Receiver for OFDM Signal with the TR Technique

Figure 2 shows block diagram of receiver for detection of OFDM signal with TR method. The received signal, r_n, can be expressed as:

$$r_n = D_n^{TR} + w_n, \quad n = 0, 1, \ldots, N-1 \qquad (11)$$

where w_n is AWGN with zero mean and a variance of $N_0/2$.

Thus, received complex symbols, $\hat{C}_k^{tr}$, are given by:

$$\hat{C}_k^{tr} = FFT\{r_n\} \qquad n, k = 0, 1, \ldots, N-1 \qquad (12)$$

If the AWGN noise term, w_n, in (11) is ignored, then (12) can be expressed as:

$$\hat{C}_k^{tr} = \sum_{n=0}^{N-1} D_n^{TR}\, e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1 \qquad (13)$$

The useful data symbols are denoted by $\hat{C}_k$, which are to be extracted from $\hat{C}_k^{tr}$ and this can be expressed as:

$$\hat{C}_k = \hat{C}_k^{tr}, \qquad k \in \mathcal{R} \qquad (14)$$

5. BER in Frequency Non-Selective Rayleigh Fading Channel

Figure 4 shows OFDM signal received over frequency non-selective Rayleigh fading channel.

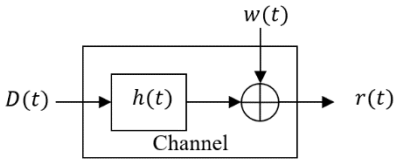


Figure 4: Model of frequency non-selective Rayleigh fading channel.

$D(t)$ in Figure 4 denotes the continuous OFDM signal of the OFDM discrete signal D_n^{TR}.

Hence, $D(t)$ is sent over the channel and therefore $r(t)$ can thus be represented as:

$$r(t) = h(t) * D(t) + w(t) \qquad 0 \le t \le T_s \qquad (15)$$

where $h(t) = \alpha(t)e^{j\varphi(t)}, 0 \le t \le T_s$ is the channel impulse response, $*$ denotes convolution, and $w(t)$ denotes AWGN.

For frequency non-selective Rayleigh fading channel, $h(t)$ can be modeled as a multiplicative distortion scaler to the transmitted signal $D(t)$. It is assumed that the channel is time-invariant and the impulse response of the channel is reasonably constant, at least during each symbol interval T_s of the transmitted signal $D(t)$. The condition of the frequency non-selective is expressed as:

$$B_s \ll B_c \text{ and } T_s \ll \sigma_\tau \qquad (16)$$

where B_s denotes the bandwidth and T_s denotes the symbol duration of the transmitted signal, whereas, B_c denotes the coherence bandwidth and the RMS delay spread is denoted by σ_τ.

Thus, the channel can be expressed as:

$$h(t) = \alpha\delta(t), \qquad 0 \le t \le T_s \qquad (17)$$

where it is assumed that α is estimated perfectly. Using (17), in (15)

$$r(t) = \alpha D(t) + w(t) \tag{18}$$

where α is the fading parameter. If α is fixed then the probability of error rate P_b expression can thus be expressed as a function of the signal-to-noise ratio (γ_b). For OFDM-QPSK system, the P_b is given by [21]:

$$P_b(\gamma_b) = \mathrm{Q}\left(\sqrt{2\gamma_b}\right) \tag{19}$$

where $\gamma_b = \alpha^2 \varepsilon_b / N_0$ and $\mathrm{Q}(x)$ is the Q function [21].

The average error probability of OFDM-QPSK system can be shown to be given by:

$$P_{b_Rayleigh} = \frac{1}{2}\left(1 - \sqrt{\frac{\bar{\gamma}_b}{\bar{\gamma}_b + 1}}\right) \tag{20}$$

6. Numerical Results

QPSK-OFDM systems PAPR performance were observed when the proposed TR scheme is used. One hundred thousand blocks of OFDM symbols are considered for simulation when $N = 128$. Both the introduced PN sequence and the phase sequence presented in [10] are examined. The reserved tones' locations are chosen randomly.

Figures 5 (a-d) show the CCDFs plots of QPSK-OFDM systems PAPR performance. The PAPR performance provided are for number of subcarriers $N = 128$ and when reserved tones are $N_{tr} = 4,\ 8, 16$, and 32. Also, these results are tabulated in Table 1.

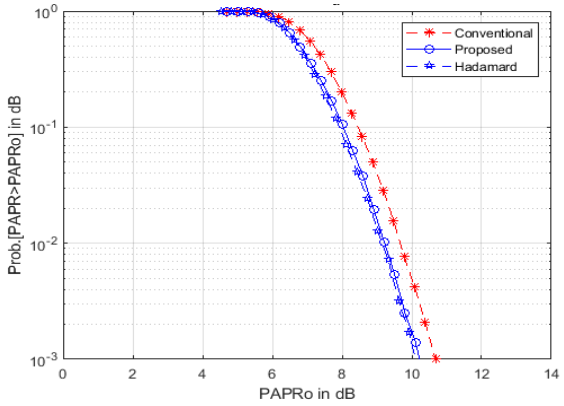


Figure 5 (a): PAPR performance of 128-subcarrier QPSK-OFDM system with 4 reserved tones

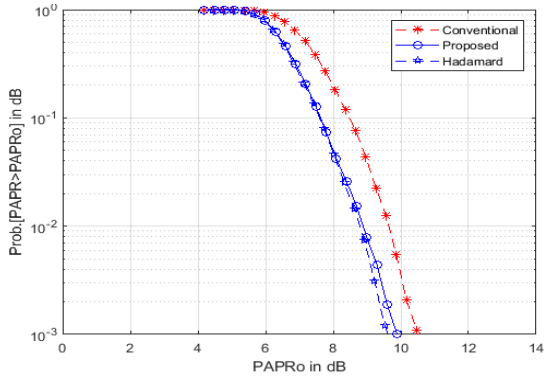


Figure 5 (b): PAPR performance of 128-subcarrier QPSK-OFDM system with 8 reserved tones

From Figures 5 (a-d) and Table 1, it is observed that the attainable PAPR performances for the two examined sequences are relatively equal. Also, it is noted that the reduction in PAPR is directly proportional to the number of reserved tones.

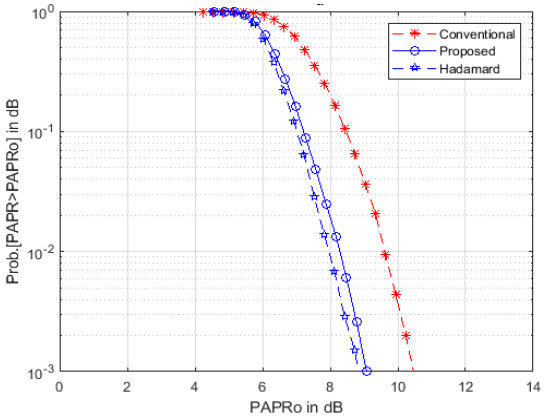


Figure 5 (c): PAPR performance of 128-subcarrier QPSK-OFDM system with 16 reserved tones

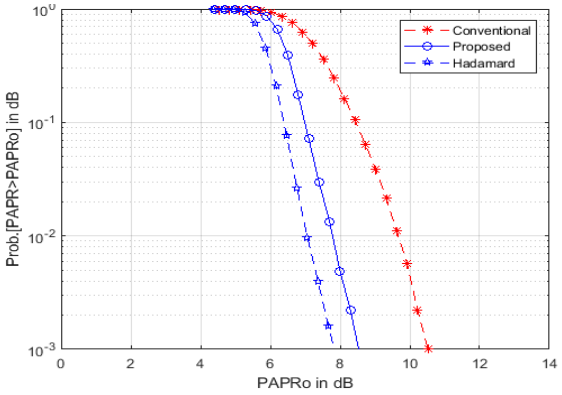


Figure 5 (d): PAPR performance of 128-subcarrier QPSK-OFDM system with 32 reserved tones

Table 1: Attainable PAPR for QPSK-OFDM system with the TR method using the proposed PN sequence and Hadamard phase sequence.

N_{tr}	Attainable PAPR Hadamard	Attainable PAPR Proposed	Attainable PAPR Conventional
4	10.286 dB	10.11 dB	10.659 dB
8	9.866 dB	10.082 dB	
16	8.992 dB	9.241 dB	
32	7.756 dB	8.457 dB	

Figures 6 and 7 illustrate the simulated BER versus SNR curves for 128-subcarrier QPSK-OFDM system with the TR technique when using the proposed PN sequence in AWGN channel and over a frequency non-selective Rayleigh fading channel respectively. Also, these results are tabulated in Table 2 and Table 3 accordingly.

Table 2: E_b/N_0 required to achieve BER= 10^{-3} for an OFDM system over the AWGN channel.

Sequence	Conventional	Hadamard	Proposed
BER	6.758 dB	6.758 dB	6.769 dB

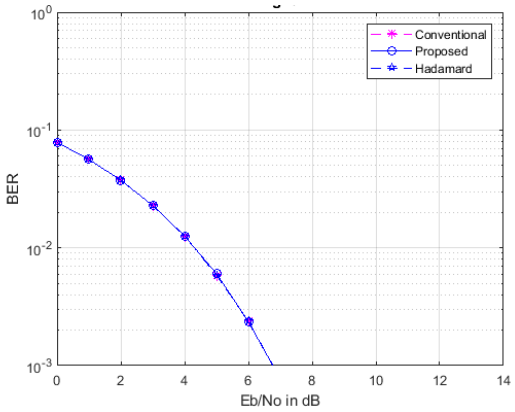


Figure. 6: BER performance of 128-subcarrier QPSK-OFDM system with proposed TR technique in AWGN channel.

Figure 7 illustrates the simulated BER versus SNR curves for 128-subcarrier QPSK-OFDM system with the proposed TR technique when using the proposed PN sequence over the frequency non-selective Rayleigh channel. Ideal channel state information (CSI) is considered.

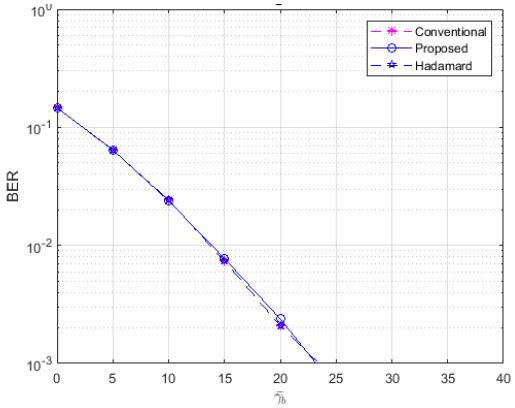


Figure 7: BER performance of 128-subcarrier QPSK-OFDM system with proposed TR technique over frequency non-selective Rayleigh fading channel.

Table 3: Average E_b/N_0 required to achieve BER= 10^{-3} OFDM system over the frequency non-selective channel.

Sequence	Conventional	Hadamard	Proposed
BER	22.936 dB	22.936 dB	23.504 dB

Figure 6 and Table 2 show that the BER performance over AWGN channel has not been degraded when the proposed PN sequence is used by the TR technique to reduce the PAPR in an OFDM system. In addition, an insignificant BER performance degradation is observed when the proposed PN sequence is used by the proposed TR method to reduce the PAPR in an OFDM system over frequency non-selective Rayleigh fading channel as shown in Figure 7 and tabulated in Table 3.

Figure 8 shows a PAPR performance comparison between QPSK-OFDM system with 16-reserved tones and QPSK-OFDM system with SLM method when the search for the optimum phase sequence is limited to 16.

Figure 8 and Table 4 illustrate that there is a 1.288 dB reduction in PAPR when TR technique is used compared to 2.497 dB when SLM technique is used. TR technique has only reserved 16-subcarriers for PAPR reduction however no SI is required at the receiver. In comparison to TR technique, SLM requires N multiplications in each iteration and it requires sending SI to the receiver.

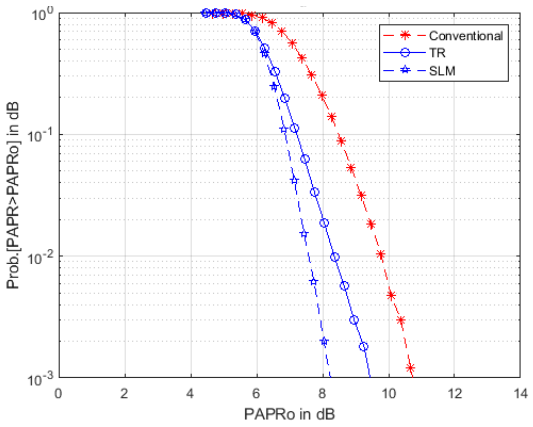


Figure 8: PAPR performance of 128-subcarrier QPSK-OFDM system with 16 reserved tones vs PAPR performance 128-subcarrier QPSK-OFDM system with SLM.

Table 4: A Comparison of the PAPR performance between 128-subcarriers QPSK-OFDM system with 16-reserved tones TR technique and 128-subcarriers QPSK-OFDM system with SLM technique

Sequence	Conventional	TR	SLM
PAPR	10.725 dB	9.437 dB	8.228 dB

7. Conclusion

In this paper, a novel low-complexity TR technique for reduction of PAPR in a QPSK-OFDM system is presented and investigated. The proposed PAPR reduction scheme is easy-to-implement and has low-complexity. Also, the proposed TR technique requires no SI to be conveyed to the receiver. The simulation results demonstrate that the PAPR performance of the proposed and conventional TR methods are the same. The BER performance of the proposed system over AWGN and frequency non-selective channel have not been compromised. When PAPR performance of the proposed TR method is compared with that of SLM technique, the SLM technique outperformed the proposed TR technique by approximately $\approx 1.2\,dB$. However, the complexity of SLM method is higher and requires sending SI to the receiver.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] U. Reimers, "Digital video broadcasting," *IEEE Communications Magazine*, 36 (6), 104-110, 1998. https://doi.org/10.1109/35.685371

[2] R. van Nee, G. Awater, M. Morikura, H. Takanashi, M. Webster and K. W. Halford, "New high-rate wireless LAN standards," *IEEE Communications Magazine*, 37(12), 82-88, 1999. https://doi.org/ 10.1109/35.809389

[3] I. Koffman and V. Roman, "Broadband wireless access solutions based on OFDM access in IEEE 802.16," *IEEE Communications Magazine*, 40(4), 96-103, 2002. https://doi.org/10.1109/35.995857

[4] W. Bolton, Y. Xiao and M. Guizani, "IEEE 802.20: mobile broadband wireless access," *IEEE Wireless Communications*, 14(1), 84-95, 2007. https://doi.org/10.1109/MWC.2007.314554

[5] Y. Tsai, G. Zhang, D. Grieco, F. Ozluturk and X. Wang, "Cell search in 3GPP long term evolution systems," *IEEE Vehi. Tech. Magazine*, 2(2), 23-29, 2007. https://doi.org/10.1109/MVT.2007.912929

[6] A. Elnashar and M. A. El-Saidny, "Looking at LTE in Practice: A Performance Analysis of the LTE System Based on Field Test Results," IEEE Vehicular Technology *Magazine*, 8(3), 81-92, 2013. https://doi.org/10.1109/MVT.2013.2268334

[7] X. Li and L. J. Cimini, "Effects of clipping and filtering on the performance of OFDM," *IEEE Communications Letters*, 2(5), 131- 133, 1998. https://doi.org/10.1109/4234.673657

[8] S. Cha, M. Park, S. Lee, K. Bang and D. Hong, "A new PAPR reduction technique for OFDM systems using advanced peak windowing method," in *IEEE Transactions on Consumer Electronics*, 54(2), 405-410, 2008. https://doi.org/10.1109/TCE.2008.4560106

[9] S. H. Müller and J. B. Huber, "A Comparison of Peak Power Reduction Schemes for OFDM," in IEEE GLOBECOM Conference, Phoenix, AZ, 1997. https://doi.org/10.1109/GLOCOM.1997.632501

[10] R. W. Bauml, R. F. H. Fischer and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Letters*, 32(22), 2056-2057, 1996. https://doi.org/10.1049/el:19961384

[11] B. S. Krongold and D. L. Jones, "PAR reduction in OFDM via active constellation extension," *IEEE Transactions on Broadcasting*, vol. 49(3), 258-268, 2003. https://doi.org/10.1109/TBC.2003.817088

[12] S. Muller and J. Huber, "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," Electronics Letters , 33(5), 368 - 369, 1997. https://doi.org/10.1049/el:19970266

[13] L. Cimini and N. Sollenberger, " Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences," IEEE Comm. Letters , 4(3), 86 - 88, 2000. https://doi.org/10.1109/4234.831033

[14] J. Tellado, "Peak to Average Power Reduction for Multicarrier Modulation," Ph.D Thesis, Stanford University, 2000.

[15] A. Jones, T. Wilkinson and S. Barton, "Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes," Electronics Letters, 30(25), 2098 - 2099, 1994. https://doi.org/10.1049/el:19941423

[16] Seung Hee Han and Jae Hong Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," in *IEEE Wireless Communications*, 12(2), 56-65, 2005. https://doi.org/10.1109/MWC.2005.1421929

[17] T. Jiang and Y. Wu, "An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals," in *IEEE Transactions on Broadcasting*, 54(2), 257-268, 2008. https://doi.org/10.1109/TBC.2008.915770

[18] Y. Rahmatallah and S. Mohan, "Peak-To-Average Power Ratio Reduction in OFDM Systems: A Survey And Taxonomy," in *IEEE Communications Surveys & Tutorials*, 15(4), 1567-1592, 2013. https://doi.org/10.1109/SURV.2013.021313.00164

[19] M. Ali, R. K. Rao and V. Parsa, "PAPR Reduction in OFDM System Using New Method for Generating Pseudo-Random Sequence for SLM Technique" in 31st Annual IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), Quebec City, Canada, 2018. https://doi.org/10.1109/CCECE.2018.8447835

[20] L. Wang and C. Tellambura, "Analysis of Clipping Noise and Tone-Reservation Algorithms for Peak Reduction in OFDM Systems," in *IEEE Transactions on Vehicular Technology*, 57(3), 1675-1694, 2008. https://doi.org/10.1109/TVT.2007.907282

[21] J. Proakis, digital communications, McGraw-Hill, 1995

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 1, 231-247 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Critical Embedded Systems Development Using Formal Methods and Statistical Reliability Metrics

Jonathan Lockhart*, Carla Purdy, Philip Wilsey

Department of Electrical Engineering and Computer Science, University of Cincinnati, 45221, USA

ARTICLEINFO

Article history:
Received: 21 December, 2018
Accepted: 09 February, 2019
Online: 23 February, 2019

Keywords:
Software Reliability
Trusted Systems
Hardware and Software Co-Design
Safety Critical Embedded Systems
Statistical Analysis

ABSTRACT

Trusted systems are becoming more integrated into everyday life. Security and reliability are at the forefront of trusted system design and are often directed at hardware-only solutions, especially for safety critical systems. This is because hardware has a well established process for achieving strong, precise, and reliable systems. These attributes have been achieved in the area of safety critical systems through the use of consistent and repeatable development processes, and a standardized metric for measuring reliability. However, due to the increase in complexity of systems and the looming end of Moore's Law, software is being incorporated more into the design of these trusted systems. Unfortunately, software typically uses agile development in modern design and uses unreliable metrics for illustrating reliability. This does not make it suitable for safety critical applications or for total system reliability in mixed hardware/software systems. Therefore, a comprehensive process of systems development needs to be utilized to allow for total system specification in the beginning and a comparable reliability metric in the end which covers software and hardware. Henceforth, we discuss an initial solution to these problems, leading to the establishment of a development process that allows for the proven correctness of a system specification via formal methods. This process also establishes a testing and error reporting process to allow software to be represented in a way that allows the application of reliability metrics similar to those used for hardware.

1 Notification of Intent

This paper is an extension of work originally presented at the 2017 National Aerospace & Electronics Conference (NAECON) [1]. In this work we illustrate the ability to utilize formal methods and automated theorem proving (ATP) to discover errors in specification development prior to implementation. A major concern addressed from feedback at NEACON 2017 is the prevention of errors in the proof environment. We expand on that work by showing that formal methods and ATPs can be used to find errors in the proof of function correctness. This allows for a system of checks and balances between specification and proof development for formal methods, verifying that both parts adhere to the original customer requirements.

This paper also extends work originally presented in 2016 at the Midwest Symposium on Circuits and Systems (MWSCAS) [2]. In that paper we addressed the use of statistical metrics to show the reliability of software over time. This process relies on the use of techniques we developed to define independent, random errors for injection into our elevator controller benchmark program. However, our presentation of those techniques was brief due to page limitations and the need to discuss the results of our testing procedure for software. To address the feedback on that work, we will be elaborating in depth on our technique for generating random, independent errors using real-world reported error results and random number generators (RNGs). This also includes a discussion of how we determine the best RNG for our needs.

*Jon Lockhart, EECS Department, 851 Woodside Drive, 812 Rhodes Hall, University of Cincinnati, Cincinnati, OH, 45221-0030, 513-556-2946 & lockhaja@mail.uc.edu

https://dx.doi.org/10.25046/aj040123

Finally, this paper extends work originally presented in 2018 at MWSCAS [3]. This was originally an expansion of our work in [2] using a different error rate and looking at possible better statistical models of reliability for software. Many of the commentators on our work suggested expanding the results to more programs and improving the results by looking into focusing the data, as some outliers had caused degradation in the results. We address both of these points by expanding the testing and analysis to five more programs, using characteristics and style similar to our original elevator controller benchmark, and then developing a process to remove outliers in the data through the use of statistical measures.

2 Introduction

Embedded systems in recent years have taken a leap in their integration with everyday society. They appear in various forms from wearables, to cellphones, to the vehicles we drive; it is important that these embedded systems function correctly. In many fields, such as medicine and aerospace, there is a strong focus on these systems being secure, reliable, and robust. These trusted and safety critical systems must work correctly as their failure could result in injury or, in the worst case, the loss of life of those that are dependent on them.

Many trusted systems encountered today utilize hardware specific embedded systems. Hardware has a tradition of reliability and safety, because of its use of industrial practices and standards. In VLSI design, developers can compare their working designs to standardized benchmarks (*e.g.*, MCNC [4], IWLS [5], and LEKO [6]) and this generates a baseline of confidence in the design. Beyond benchmarks, tool designers have incorporated Model Checking [7] into their development suites, allowing for reliability checking against industry determined "golden models" [8–11]. These techniques are invaluable in generating the statistical models of reliability for hardware systems.

Unfortunately, the constraints of Moore's Law [12, 13] and the increased complexity of system function have required a change in the design approach of critical embedded systems. Some industry practice has attempted to maximize the resources available for a given space [14–17], but these techniques are fairly new, and VLSI tools and techniques are lagging behind in allowing for these techniques, slowing adoption. Other researchers are looking at maximizing the capabilities of current hardware by focusing on implementing machine learning in embedded systems [18]. However, this is difficult to achieve for safety critical systems with the large overhead that comes with high precision computing with machine learning [19].

Hence, software is an attractive alternative for meeting the demand for complex, real-time critical systems. Several recent disasters (*e.g.* [20–22]) have, however, indicated that the current process of testing to exhaustion and representing software reliability as defects per thousand lines of code (ELOC) [23] is not sufficient. This is especially true when taking into account the fact that there is no industrial standard for counting "lines of code" and factors such as size play a role in misrepresenting results [24]. Software metrics similar to hardware would be ideal, and theoretical work in this area has shown software should be able to be modeled with statistics [25–30]. Government agencies have also released standards of development [31, 32] to provide a framework for illustrating reliability effectively. However, there is no specific development process outlined for achieving either of these improvements.

The work presented here is a complete development process that we propose for use in addressing the challenges described above. This development process allows for the establishment of a suite of benchmark programs, defining a baseline similar to hardware which other trusted systems utilizing software can be compared to. This suite is open ended and allows for the addition of more programs in the future. To accomplish this goal, we will address the area of specification design through the use of formal methods, and show how errors can be eliminated early in development with the use of ATP. This work is an expansion of work already presented in [1, 33]. We will also illustrate the development, use, and statistical results for reliable software in safety critical applications. This approach has been briefly presented previously in [2] and improved in [3]. We expand upon those findings to complete the suite of small embedded programs which establishes the initial set of well-defined benchmarks for software reliability and a baseline for software testing.

3 Related Work

There have been several groups researching the use of formal methods in the recent literature. This work has targeted two main areas. The first is the use of model checking with software, and the second is the improvement of ATPs. One area of use for model checking is for reverse engineering a system [34]. Since automated implementation is not available yet for specifications written in formal methods, the reverse engineering process uses the implementation to move back to a specification, verifying it is correct with model checking, and then verifying the original specification and reverse engineering specification are correct with respect to each other. Another area is related to the development of a tool called Evidential Tool Bus (ETB) that combines both model checking and ATPs into a single environment to allow better verification of specifications written in formal methods [35]. Finally, researchers are looking at the inclusion of model checking into development tools to use in verifying that object oriented programs are correctly implemented [36]. This process is similar to the current tools for model checking hardware designs provided by Synopsys [7].

The research in improving ATPs has involved making them more inclusive for developers. Currently

systems such as ProofPower [37], although powerful, require the user to know the formal method and ATP language for proof verification. Some work has been done to provide a more mathematical environment to work in, replacing the intermediate ATP language, *i.e* SML, with mathematical notation. Some of this work includes the use of superposition calculus [38], modal logic [39], and standard calculus [40]. Other researchers have looked at changing the language of formal methods and ATPs together, working to use traditional programming languages from the beginning of specification design, and this work has focused primarily on the language LISP [41]. These changes would require a change in the current specification writing practices with formal methods [42, 43].

Current work in the field of software development has focused on improving the metrics of throughput and performance for large scale, big data [44] applications. This work includes generating heuristics for quality assurance (QA) [45] with probabilistic analysis, and attempting to predict where in the development life cycle a particular project will have the most errors [46, 47]. This allows developers to focus their efforts in those areas to mitigate the time and cost for a given project. Machine learning is assisting in this field, where it is used to predict the optimal release time of a given system based on the metrics of testing, cost, and errors produced [48]. Many of these techniques are being applied specifically to projects that rely on cloud computing [48] and open source software [49]. They are not being used to directly target the component level reliability which is required for safety critical systems and those relying on software in an embedded environment.

4 Background Knowledge

Formal methods are supported by many languages, such as Z [50] and VDM [51], that share in common a reliance on basic mathematical principles. These principles include axiomatic notation [52, 53], predicate calculus [54, 55], and set theory [56, 57]. When combined with the formal methods syntax, these principles allow us to develop specifications describing the functionality of a system [22,58]. These principles are often taught in engineering course work, and formal methods often have manuals describing how the language operates [59], similar to a conventional programming language such as C or Java.

One important aspect, however, is the mechanisms through which formal methods have traditionally been used to show that a given method in a specification is correct. The traditional method is to prove a specification is correct through a process of refinement. This process requires "refining" a particular part of a specification written in formal methods to an implementation method, such as analog circuitry or Java, for a particular portion of the specification. This decision to refine to a particular implementation is required ahead of time so there is a goal to achieve for the refinement. Later changes to the implementation require that a new proof via refinement be performed. Understanding this process and why our work has moved in a more modern direction is important and we illustrate this process in the following example.

The example we are working with will utilize an elevator controller benchmark as originally described in [33] and referenced in Section 5. The elevator controller discussed here has requirements presented in Figure 8. For the refinement process, we will use Z for the formal specification language and C as the implementation target. Note that any formal method and implementation medium can be utilized for refinement as long as the characteristics of each are well understood and the proof can be completed. In contrast to the original requirements of [33], some additional constraints need to be in place before the refinement can proceed. In particular, the constraints are:

- The model will illustrate the action of users in the system; specifically we will model the entry of passengers onto the elevator
- People are unique individuals
- Entry into the elevator is sequential
- The elevator can move with or without people
- There is no max capacity
- The system contains one elevator

With these extra constraints on the system defined, the refinement proof can begin. The first step is to develop a definition of the possible group of users for the elevator. Since we used the wording in the constraints, we shall refer to these users as "People." We use Z to define "People" as a "Set," because sets are flexible and convenient for developer definitions in Z. Figure 1 illustrates the definition of "People" in Z.

Figure 1: Z definition of the set "People"

Now that the generic definition of "People" has been established, we need to define a subset of "People" who will be using the elevator. To ensure that the specification is as open as possible we will define the subset of "People" who are riding the elevator via a power set. This way any combination of members of "People" can become "riders." Figure 2 captures the creation of this power set using the Z construct known as a state in ProofPower [59].

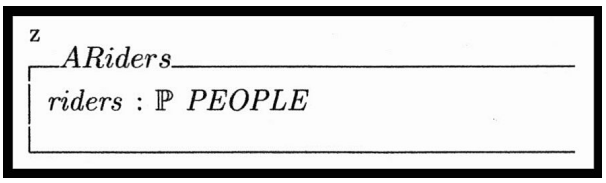


Figure 2: Generic state sefinition for the power set "Riders"

The final part of the initial definition is to describe the act of getting on the elevator. In Z this can be done using a function declaration, which will include a change in the state of the system and a member of the set "People" becoming a member of the power set "riders." Figure 3 shows the generic definition for this function.

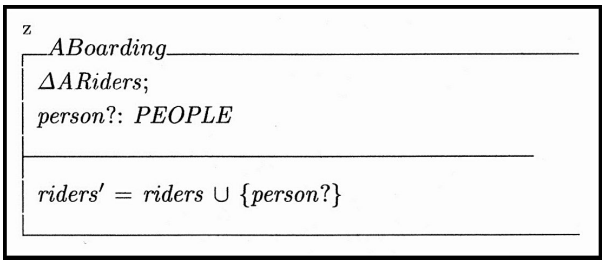


Figure 3: The generic function definition for boarding the elevator

Now we need to complete the same process but with the implementation medium we have decided on. As mentioned, we have chosen C for this task. The first step is to define the state of the system as it would be represented in C. We will reuse the same set "People" we defined previously for this task. In a programming language such as C, sets of items are usually stored in an array or a linked list. This is what we would use to store the members of the set "People" who board the elevator. However, Z does not have these constructs directly, though it does have a structure that an array or linked list can be modeled in, and that is a sequence. A sequence is a good comparison as a set is equal to the range of a sequence, as shown in (1).

$$s = ran\, ss \tag{1}$$

In this equation *s* represents the set, *ran* is the notation for range, and *ss* represents a sequence. Now we construct the state of the system with Z for C using the sequence term *seq*, and generating a new, more specific group called "riderss." Figure 4 shows the construction of this state.

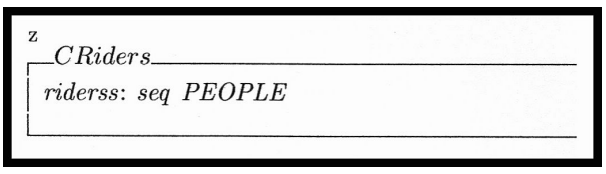


Figure 4: State for "Riderss" in C using Z

Finally, we complete this portion by describing the action of a new rider boarding the elevator in C. Since we are dealing with a sequence, the mathematical notation we will use is a concatenation. Figure 5 shows the state for C as defined in Z.

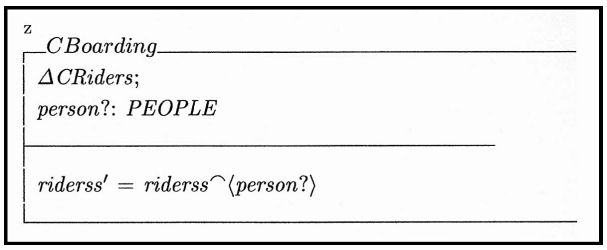


Figure 5: C function definition for boarding the elevator

With the set "People" we have now defined two versions of the system for boarding the elevator, a generic one which describes the basic functionality and a more specific one constrained by the properties of our implementation medium C. This gives us a starting and ending place. To bridge the gap between the two and what they describe, we need to develop a proof to show that the generic function can become the specific function. To start we build a proof implication, or goal, illustrating this transformation. This will be done using Equation 1 for some replacements for *riders* and *riders'* from the generic definition in Figure 3. Figure 6 gives the initial implication, showing the end requirement for the C definition from Figure 4 and using the previous set to sequence rule to rewrite *riders* and *riders'* in terms of the range for *riderss* and *riderss'*, respectively.

To complete the proof implication, we want to drive one side or the other to an absolute *true* which means everything on the other side must hold as defined. Since the outcome only has one statement, this is the easier of the sides to work on. Figure 7 shows the seven step proof indicating the implication holds and the refinement is *true*.

Using set theory and rewrites of variables taken from the specifications, we are able to show that the generic specification can be refined to the specification for C. This was a simple example and even then it took seven steps to complete, along with a strong knowledge of the constructs of Z, C, and set theory. To attempt to complete a definition for a more complex system and multiple functions would be very time consuming and may cost more than just using traditional testing methods due to the time involved. Furthermore, this process does not define the function in a useful manner that would assist with the conversion to an implementation, but rather shows the theoretical correctness of the system definitions for a particular implementation medium. Even when using an ATP to perform the refinement steps, the final implementation may not be correct in a functional sense, as the definitions are mathematical, not practical. Hence our work [1, 33] described in Section 5 and extended in Section 6 shows the advantages of utilizing a functional specification and proving that correct. It also provides the flexibility of being implementation neutral so a new proof does not need to be constructed if the implementation

$$riderss' = riderss \frown \langle person? \rangle \land riders = \text{ran } riderss \land riders' = \text{ran } riderss' \Rightarrow riders' = riders \cup \{person?\}$$

Figure 6: Implication definition for showing the refinement steps

$$riderss' = riderss \frown \langle person? \rangle \land riders = \text{ran } riderss \land riders' = \text{ran } riderss' \Rightarrow riders' = riders \cup \{person?\}$$

$$\begin{aligned}
&1 \Leftrightarrow riders' = riders \cup \{person?\} \\
&2 \Leftrightarrow \text{ran } riderss' = riders \cup \{person?\} \\
&3 \Leftrightarrow \text{ran } (riderss \frown \langle person? \rangle) = riders \cup \{person?\} \\
&4 \Leftrightarrow \text{ran } riderss \cup \text{ran } \langle person? \rangle = riders \cup \{person?\} && [\text{Law: ran } (s \frown t) = \text{ran } s \cup \text{ran } t] \\
&5 \Leftrightarrow \text{ran } riderss \cup \{person?\} = riders \cup \{person?\} && [\text{Law: ran } \langle x? \rangle = \{x?\}] \\
&6 \Leftrightarrow riders \cup \{person?\} = riders \cup \{person?\} \\
&7 \Leftrightarrow true
\end{aligned}$$

Figure 7: Proof showing the refinement of the specifications using Z

medium changes.

One final area of background work that we utilize for our error analysis is the theory that software errors can be modeled as a Non-Homogeneous Poisson Process (NHPP). A NHPP is defined well in the sources [60–63]. The important factor for our work is that software errors were shown to meet the requirements of a NHPP in [29], which originally had been theorized in [25, 26, 30], and shown to fit statistical models in [28]. The takeaway is that software errors have been shown to fit NHPP, implying that these errors occur at random intervals and that they are independent of one another. This allows us to define our error generation and testing procedure, discussed in Sections 7 and 8 and in the results presented in [2, 3].

5 Specification Development with Formal Methods

As shown in Section 4, an understanding of refinement is important for developing specifications and proofs for showing the system design is correct. However, this traditional methodology limits the scope of the system to a particular implementation method. The implementation chosen for the proof reduction my not be the best for a given system process later in development. If a new implementation is chosen, then a new proof will need to be created to drive the specification to that implementation, resulting in the development of new state diagrams and another refinement. This correction in the design would be costly in the time required for the refinement.

In previous work [33], we utilized a simple elevator controller, in the style of Moore and Mealy Machines, to show that an implementation independent specification could be developed and proven correct without the need for traditional refinement. This was done using "Z" [50] and the Automated Theorem Prover (ATP) "ProofPower" [37]. Figure 8 shows an excerpt from [33] with the initial customer requirements for the simple elevator controller.

- Must know which direction it is going.
- Can only move up and down.
- Has only two inputs: the starting floor and ending floor.
- Must stop once it has reached the desired floor.
- Must have an emergency state to shut down the elevator.

Figure 8: Customer requirements for our elevator controller

From these requirements we are able to build a complete specification and proof illustrating our formal methods design process [33]. This added step in development does add time to the process, just like refinement, but there are strong benefits to its use over refinement. This includes focusing the functionality of the system, making it less likely errors will occur during the implementation step [33], since the specification is designed in a similar style to function definitions in software. This also eases the transition when building an implementation from customer requirements. Another benefit is finding development errors prior to implementation [1], which is usually a cost saving measure compared to finding errors later in development. There is also the added advantage of not having to refine the specification to a particular implementation, while maintaining the advantage of proving the system correct.

The error now is in the proof goal set by the independent tester, causing the goal statement to be $calling_floor? \geq starting_floor?$, which is shown in Figure 10.

6 Using Formal Methods to Eliminate Design Errors

An important advantage to using Formal Methods is to allow for design flaws to be determined prior to the

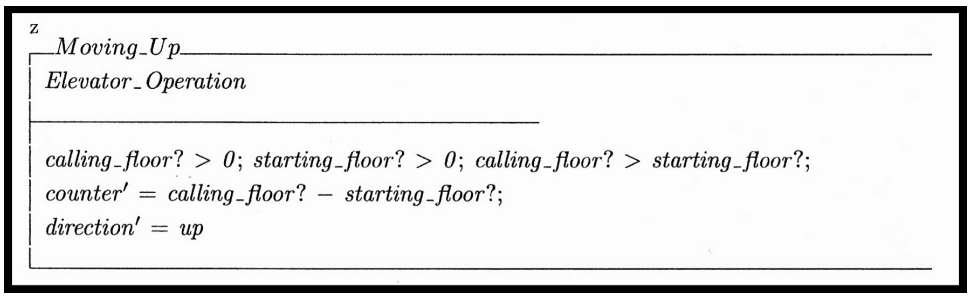


Figure 9: "Moving Up" function represented in Z

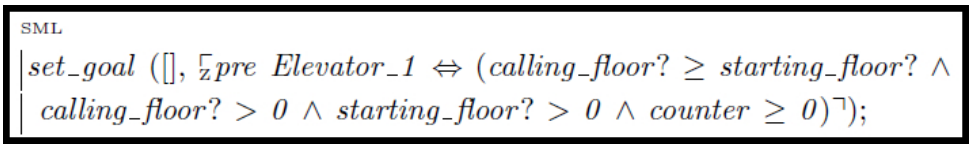


Figure 10: Proof goal for "Moving Up" with an incorrect math statement

system reaching implementation. Often a customer's requirements may have conflicting or impractical goals which cannot be determined until they are built into the system. Formal Methods proves those requirements are sound prior to implementation, decreasing the time of development and the cost to remove errors later in development. However, human error can still plague the development life cycle.

Human error in development can come in many forms, but one such error in software design comes from the developer accidentally injecting an error into the system. For example, misspelling a variable name so it matches a similar variable could cause a drastic change in the functionality of the system. Another example is changing an addition to a subtraction, which could be accomplished with an errant key stroke. To combat these errors, independent verification is used in testing, helping to eliminate the bias a developer may put unknowingly into their own tests. This concept works equally well with Formal Methods.

In our work [1] we previously showed that independent verification during the proof of a specification enabled errors to be found and eliminated prior to implementation. Feedback on this process noted that the error could be propagated from the specification to the proof, causing it not to be caught. We emphasize that specification and proof development should be completed by separate individuals, allowing for independent verification. It is unlikely that two developers would make the same mistake by misinterpreting the customer requirements. This does however bring up a separate issue, *i.e*, that the mistake could be with the proof and not with the specification. Would the proof environment be able to locate an error in the proof statement, given a correct specification? This is the question that was posed, and this is what we have set out to determine.

In [1] we utilized the "Moving Up" function as test and that is the function utilized here for consistency. Our work in [1] showed that a slight change in the specification over the customer requirements could cause a dramatic change in how the system functioned. In this case the original requirements, as shown in Figure 8, require that, if the elevator moves up, the floor it is going to has to be a higher value than the floor the elevator starts on. This is represented in the specification as $calling_floor? \geq starting_floor?$, where $\geq$ has replaced the correct term $>$.

Now, consider the reverse situation. The original requirements have remained the same from Figure 8, and now the specification for "Moving Up" is correct as it is in [33], shown in Figure 9.

With this mistake in the proof, the proof begins with a goal that is less strict than the specification, in that "greater than or equal to" is more inclusive than "greater than." Much like our previous use of ProofPower in [1, 33], we need to drive the proof to completion, attempting to resolve all goals defined in Figure 10 via a series of commands issued to the ATP. Figure 11 shows the current steps we will use to try to prove the specification is correct.

Step 10 in Figure 11 is where the proof fails. Figure 12 shows the output from ProofPower when the exception is thrown and the proof can not be completed. ProofPower is reporting that the original goal as stated can not be proven given the assumptions and constraints on the proof, which were set in the original specification.

Unlike the original error recognition test in [1], the semantics error takes longer to find when the error is in the proof goal; nevertheless the error was found, even when it was within the proof goal definition and was weaker than the original specification. Overall, the exception shows that there is an issue with the system, either with the specification or the proof. At this point it would be up to the proof writer and the developer to review their respective sections with regards to the customer requirements and to determine where the error occurred.

Finding errors early in development, as in this example or previously in [1], is invaluable to the development process. This allows you to find semantic errors

```
1  a (rewrite_tac [Elevator_1, Moving_Up, Elevator_Operation, Elevator_State]);
2  a (pure_rewrite_tac [z_get_spec ⌜(_≤_)⌝]);
3  a (rewrite_tac []);
4  a (REPEAT z_strip_tac);
5  a (asm_rewrite_tac []);
6  a (PC_T1 "z_lin_arith" asm_prove_tac []);
7  a (z_∃_tac ⌜(direction' ≙ up, counter' ≙ calling_floor? + ~ starting_floor?)⌝
     THEN rewrite_tac []);
8  a (asm_rewrite_tac [z_get_spec ⌜DIRECTIONS⌝]);
9  a (PC_T1 "z_library_ext" asm_rewrite_tac []);
10 a (PC_T1 "z_lin_arith" asm_prove_tac []);
```

Figure 11: The steps necessary to complete the proof to show the goal is correct

Table 1: PSP program line count and error results

Language	Program	Total Lines of Code	Total New Lines of Code	Number of Errors
C++	1	82	82	12
	2	233	233	31
	3	463	263	24
	4	236	129	15
	5	178	104	19
	6	568	299	40
	7	678	86	20
	8	458	298	47
	9	824	258	35
	10	1202	379	62
TOTAL		4922	2131	305
Eiffel	1	70	N/A	6
	2	82	N/A	17
	3	214	N/A	19
	4	220	N/A	8
	5	182	N/A	13
	6	396	N/A	28
	7	286	N/A	11
	8	444	N/A	20
	9	784	N/A	23
	10	766	N/A	36
TOTAL		3444	N/A	181

that are caused by developer mistakes prior to those errors being propagated into the implementation. If this were to occur, there is no guarantee that the error would be caught in the testing phase, and thus it could require a change to the system after release, costing extra time and money. Finding errors earlier in the development process is usually more cost effective, and utilizing a well developed ATP like ProofPower should help eliminate development errors prior to implementation.

7 Developing Repeatable Errors from Real World PSP Results

Before moving into implementation and testing, it is crucial to have repeatable results. In our case, this means not repeatedly injecting the exact same system errors. This is important as the premise of some of our work goes back to the work done by Yamada [29, 64], as explained in Section 4. Our errors need to be random and independent of one another for them to fit into the model of an NHPP. This will not be the case if the same errors are used for every evaluation of our procedure. To that end we turned to the work of Victor Putz [65] who created examples using PSP [66] to illustrate its ability to assist a developer in becoming more proficient in software development. Through the posted results we are able to catalog the findings, developing metrics to allow us to generate error rates and determine errors for replication in our test system.

Putz performed a self-exploratory study through PSP, publishing the information for each step in the training process, along with the code he developed, his

Output Continued

```
...
:) a (PC_T1 "z_lin_arith" asm_prove_tac []);
Cannot derive a contradiction from the following assumptions using the linear
 arithmetic proof procedure:
⌜z_ℤ $"calling_floor?" + ∼ (z_ℤ $"starting_floor?") ≤ ℕℤ 0⌝
⌜∼ (z_ℤ counter) ≤ ℕℤ 0⌝
⌜∼ (z_ℤ $"starting_floor?") ≤ ∼ (ℕℤ 1)⌝
⌜∼ (z_ℤ $"calling_floor?") ≤ ∼ (ℕℤ 1)⌝
⌜∼ (z_ℤ $"calling_floor?") + z_ℤ $"starting_floor?" ≤ ℕℤ 0⌝
Assigning variables to terms as follows:
X1 = ⌜z_ℤ counter⌝
X2 = ⌜z_ℤ $"starting_floor?"⌝
X3 = ⌜z_ℤ $"calling_floor?"⌝
Gives the satisfiable system:
1: −X2 +X3 ≤ 0
2: −X1 ≤ 0
3: −X2 ≤ −1
4: −X3 ≤ −1
5: +X2 −X3 ≤ 0
Exception Fail * Could not prove theorem with conclusion ⌜ℤ(0 ≤ calling_floor? ∧
 0 ≤ starting_floor?) ∧ 0 ≤ calling_floor? + ∼ (starting_floor? + 1)⌝ (* System
 is satisfiable [ℤ_lin_arith_tac.82110] *) [z_lin_arith_prove_tac.82200] *
 raised
```

Figure 12: "ProofPower" output from failing step ten in the reduction process

thoughts, issues, experience, and results [67]. He used a chart system to categorize each error that he made for each program, and an example of this report can be found in [68].

There are altogether ten programs developed through the PSP training, and many of the previous programs build on one another. In Putz's analysis of the training, he went through the process using two different languages, C++ and Eiffel. Each language had its own unique set of code [67], and statistics for each based on the program developed. C++ did have better reporting as compared to Eiffel, but regardless, based on the information and programs reported for each chapter, we were able to collect data from the study. Table 1 shows the results we collected from what was presented by Putz.

Between the two programs, C++ had total lines developed and total new lines added (for programs that were expanded upon), whereas Eiffel just had the lines for each program developed. Despite this we are able to determine several error rates, two from C++ and one from Eiffel. Eiffel's error rate was calculated by taking the total number of errors committed and dividing it by the number of developed lines. This rate is 5.26%. C++ programs yielded two error rates, one from total developed lines and one from total new lines added, with both being divided into the total number of errors committed in the ten programs. This gives us two very different rates for C++, with one being below ten errors per lines of code (ELOC) and one over. The high error rate for C++ is 14.31% and the low error rate is 6.2%.

Now that we have rates to use to inject errors into a system per one hundred lines of code, we next need to establish what error groups were reported. From the result tables, like the one in [68], we can see that each error is coded, and a text description of the error and/or its resolution is provided. We refer to the chart Putz posted on his site to get the categories and their meaning [69]. Using this information we are able to generate a list of the errors that appeared during his independent study of PSP. From the error categories and definitions, along with examining the code, we were able to break the errors into groups. We then determined the errors made for each type in each program for C and Eiffel. Table 2 lists the errors of each of the types for each program along with their totals.

Based on the categories of errors presented, we determined what errors could and could not be replicated using automation. The errors categorized as "Algorithm Alteration," "Missing Algorithm," "Specification Error," or "Unknown" are more overarching, sometimes requiring a redesign of a whole algorithm, and would be too difficult to reproduce automatically. Those errors should occur anyway in the implementation and testing process for a developer, so our programs should have representatives of these errors included in them naturally. The remaining errors could be replicated via selection, which is addressed in Section 8, because they affect a specific point in the code, requiring up to only a few changes to correct. Table 3

shows the final selection of errors chosen for replication along with the rate at which they occur during the PSP process. These rates were calculated by taking the total found for a specific error and dividing it by the total number of errors for error types we can replicate.

Table 2: Total erors made by type for C++ and Eiffel

Error Type	C++	Eiffel	Total
Spelling	29	26	55
Used Wring Type	52	26	78
Missing Header Information	21	6	27
Incorrect Math Operation	28	33	61
Incorrect String Operation	11	10	21
Error Handling	13	2	15
Passing Parameters	12	3	15
Conditional Error	10	11	21
Incorrect Method	16	16	32
Missing Block/End Line Character	30	12	42
Algorithm Alteration	31	18	49
Missing Algorithm	37	15	52
Specification Error	11	2	13
Unknown	4	1	5

Table 3: Percentage of Errors Made by Type

Error Type	Percentage
Spelling	14.99
Used Wrong Type	21.25
Missing Header Information	7.36
Incorrect Math Operation	16.62
Incorrect String Operation	5.72
Error Handling	4.09
Passing Parameters	4.09
Conditional Error	5.72
Incorrect Method	8.72
Missing Block/End Line Character	11.44

From the diagnosis of Putz's results we have been able to develop error rates to use for injecting errors into our system, and rates for error types that we can replicate. This will be valuable and completes the first step, namely the need for independent errors. The next step will be formulating a way to determine which errors will occur, how many will occur, and where each will be injected into the system.

8 Generating Random Errors for Insertion Into a System

Previously in [2] we briefly discussed an analysis to determine a good RNG for use in replicating random errors. A critique from the results presented was the need to test the RNGs at all. The most critical response to this is that the errors being generated need to be random and independent of one another such that the errors will fit the aforementioned NHPP. Unfortunately, RNGs for general computing, such as `rand()` that comes with C/C++, are only pseudo-random. This means that there is a pattern to the numbers being generated and it could be predicted. If the values are predictable then the numbers generated are not random and therefore will not fit NHPP.

Tests have been generated over the years to determine the randomness of an RNG, *e.g.*, by Knuth [70,71] and Marsaglia [72]. However, these tests are singular and require the tester to input parameters to constrain the test. This allows an RNG developer to design their generator to defeat one or two of the tests chosen to show randomness. Fortunately L'Ecuyer has provided a solution to this issue with the `TestU01` package [73]. This C package combines multiple tests from multiple authors, and has been calibrated in such a way that each test is given the correct parameters for the level of testing being performed [74]. For this work we will be utilizing the test suite L'Ecuyer calls `SmallCrush` as it is described as sufficient for determining if an RNG is suitable for general computation [73]. All inputs and test generations for `TestU01` have been determined by L'Ecuyer using statistical analysis [74], and in the case of `SmallCrush` the total number of values generated for each test is approximately 5 million values, which we validated through testing.

Using `SmallCrush` we can determine an RNG that can be considered sufficiently random and allow for meeting the requirements for NHPP. `Rand()` was decided upon as a baseline for utilizing `SmallCrush` and generating results to see what tests it would pass. `Rand()` was also chosen as the baseline since it is the default RNG included with the C/C++ library. Testing with `SmallCrush` was completed three times for consistency, on integer and decimal values, and was initially seeded at the start of each test with the current time to prevent starting at the same value for each test. Figure 13 shows the an example of the output from `SmallCrush` using `rand()` to generate integer values, and Figure 14 shows the final results for `rand()` generating integer values.

The results show that `rand()` for C/C++ fails twelve of the total fifteen tests (passing only three). This is consistent for the next two runs, failing the same twelve tests each time. Now with this baseline we can move on to see if we can find a better RNG, needing one that can pass all fifteen tests to be considered "random."

We want to minimize the development of tools and inclusion of extra libraries, so finding a generator that

can be random for both integer and decimal values is optimal. Researching the literature, one in particular that appears to meet these requirements is PCG32 by O'Neal [75]. This RNG has a small extended footprint with its included libraries, and the results show that PCG32 has already passed SmallCrush. We will independently re-evaluate these claims. Other RNG designs were considered, such as SBoNG [76], but they have already been shown not to be "random" [77] via SmallCrush.

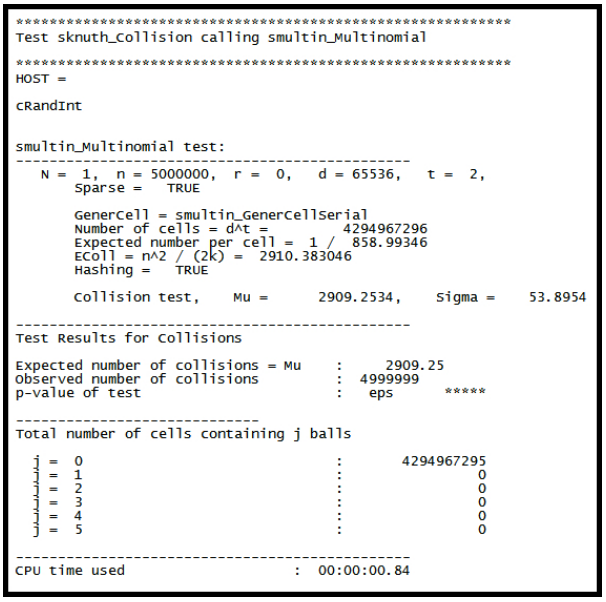


Figure 13: Output from SmallCrush testing rand() RNG

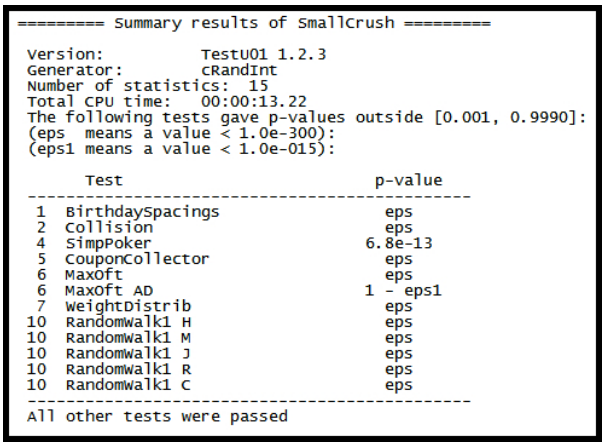


Figure 14: Final results for rand() when tested with SmallCrush

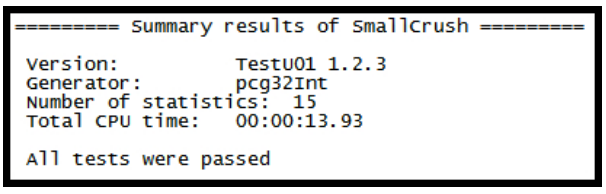


Figure 15: Final results for PCG32 when tested with SmallCrush

PCG32 was tested three times each for consistency, just like previously with rand(). Figure 15 shows the results for one run of PCG32 generating integer values. All three runs through SmallCrush resulted in the same outcome.

The results from this testing are impressive. The results show that PCG32 performs as claimed [75], passing each of the tests for all three runs. It has proven to be sufficiently random for general computation and allows us to generate the random values necessary for compliance with NHPP.

9 Random Error Generation

Before we can evaluate our testing procedure, we need to generate random errors for the system using the error rates we determined in Section 7. This is subdivided into three parts:

1. Count the lines of code for each benchmark program in a consistent manner
2. Determine the possible injection points for each error we will replicate; this will need to be done for each benchmark program
3. Randomly select the number of errors to occur, which errors will occur, and where in the program using the derived data from the PSP analysis

Item 1 is the first to be addressed. An automated method for determining the number of lines for each program is critical for the count to be consistent and accurate for each benchmark program. Unfortunately there is no industry standard for counting lines. Many reports have proprietary methods of counting lines, such as compiled compared to written, or counting spaces versus not counting spaces. This is typically used to help improve the error rate for ELOC representation, as mentioned earlier. In our case, in reviewing the lines reported by Putz, and reviewing our own benchmark programs, it was determined that the fairest and most consistent measure would be to count the lines of written code that would be compiled. This means only complete lines would be looked at, and unnecessary spaces and trialing lines with braces would need to be eliminated. This would compress the written code down into lines to be compiled while allowing each required statement to be on its own line. Figure 16 shows an excerpt of code from our elevator controller benchmark after it has been compressed by the line counter.

The results show the extent of the function of our counter, removing unnecessary space for the functions of the code (usually they are for clarification), and comments (which are not necessary for function and skipped by the compiler). Also, braces have been moved to the last complete line that requires them (meaning that some lines have multiple braces from nesting). Overall, this makes for a compact and fair result. Each of our six benchmark programs was processed through the counter. Table 4 shows the final lines counted for each program.

```
#include <stdio.h>
#include <stdbool.h>
enum DIRECTIONS = {up, down, stop, off};
int main() {
        int floors = [1, 2, 3, 4, 5];
        int startingFloor = 1;
        int callingFloor = 1;
        DIRECTIONS direction = stop;
        bool failure = false;
        char text(20);
        print("Starting Up Elevator...\n");
        sleep(1.5);
        print("Opening Doors...\n");
        sleep(2);
        print("Initial Passengers Boarding...\n");
        sleep(5);
        do {
                fputs("Enter Destination: ", stdout);
                fflush(stdout);
                if(fgets(text, sizeof text, stdin)) {
                        if (sscanf(text, "%d", &callingFloor) == 1) {
                                printf("Number Entered: %d\n", callingFloor);
                                if ((callingFloor <= 0 || callingFloor > 5) && callingFloor != -1) {
                                        failure = true;
                                        emergencyShutdown();}
                                else if (callingFloor = -1) {}
                                else {
                                        direction = determineDirection(startingFloor, callingFloor, direction);
                                        if (direcation == up) {
                                                movingUp(startingFloor, callingFloor);}
                                        else {
                                                movingDown(startingFloor, callingFloor);}}}}}
        while (callingFloor != -1 || !failure);
        if (faliure) {
                print("Maintenance Password Entered Properly!\n");
                print("Elevator can now be Shutdown for Service!\n");
                getchar();
                print("Shutting Down!\n");
                sleep(5);}
        else {
                print("Program is over...\n");
                print("Returning Elevator to Bottom Floor...\n");
                movingDown(startingFloor, 1);
                print("Please Press Enter to End Program!\n");
                getchar();
                print("Program Will End in 3 Seconds...\n");
                sleep(3);}}
```

Figure 16: Elevator controller code after being parsed by our line counter

Table 4: The line counter results

Program	Total Lines Counter
Elevator Controller	96
Automatic Door Controller	62
Security System	89
Streetlight	83
Tollbooth	54
Vending Machine	61

The next step is Item 2. This proved difficult to automate, as the C library is vast, and developing a parser to incorporate all the possible injection sites for each of the error types could not be completed. In the future we plan to automate this process, but for now this was completed by hand through scanning the code for each error type and marking an injection site with a value, and then counting the number of places an error could be injected. Some errors had overlap, so care will need to be taken to incorporate both errors if there is overlap when injection sites are randomly selected.

Finally, Item 3 is completed. Our program allows us to input the number of lines determined from the line counter program, and the number of places each error can be injected. The error selector is hard coded with the C++ high and low error rates determined in Section 7, as those will not change. The algorithm uses the rate, which is a percentage, the average errors seen per one hundred lines of code, and then determines the number of errors that should be injected into the current program. For example, if the program had two hundred lines of code, then, based on the high error rate, 28.62 errors should be injected into this program. Since we can not generate fractional errors, the system rounds up to the nearest integer for rates $x.y$ where $.y \geq .5$ and rounds down to the nearest integer for rates $x.y$ where $.y < .5$. In a similar fashion, using the total number of injected errors just calculated, the program calculates the number of errors of each type to generate. This is done by having PCG32 generate a decimal value z within the range $0 \leq z \leq 100$. The previous error percentages are now on a scale from 0 to 100, with subdivisions ending with their rate plus the previous rate. Table 5 shows the range for each error.

Table 5: Repeatable error ranges for random generation

Error Type	Decimal Range
Spelling	$0 \leq z \leq 0.1499$
Used Wrong Type	$0.1499 < z \leq 0.3624$
Missing Header Info	$0.3624 < z \leq 0.436$
Incorrect Math Operation	$0.436 < z \leq 0.6022$
Incorrect String Operation	$0.6022 < z \leq 0.6594$
Error Handling	$0.6594 < z \leq 0.7003$
Passing Parameters	$0.7003 < z \leq 0.7412$
Conditional Error	$0.7412 < z \leq 0.7984$
Incorrect Method	$0.7984 < z \leq 0.8856$
Missing Block/End of Line Characters	$0.8856 < z \leq 1$

The system repeats the process of using PCG32 to generate a value and comparing it to the scale in Table 5 until all errors for the system have been generated. Table 6 shows an example of the number of errors of each type selected randomly for a single run of the program for the Elevator Controller for the high error rate.

Finally, the last step is to determine the locations for the errors to be placed for the number of each error selected. These locations are also randomly selected using PCG32. For example, Table 6 shows that, for the Elevator Controller, 3 spelling errors are to be made. Using the selected number of errors by type, and the total number of places in the code that error can occur,

the system randomly selects the locations for injection into the elevator code. The system is configured not to select the same injection site more than once.

Table 6: Error selection for the first elevator test run

Error Type	Total Generated
Spelling	3
Used Wrong Type	2
Missing Header Info	1
Incorrect Math Operation	1
Incorrect String Operation	1
Error Handling	1
Passing Parameters	0
Conditional Error	2
Incorrect Method	0
Missing Block/End of Line Characters	3

For multiple testing runs, the program is run multiple times to generate a unique set of error injections for each run of a program, with no two runs being alike. For our testing, each program is going to be put through the testing procedure ten times, which requires ten selections of error types and locations to be processed. This process is then repeated for the low error rate as well, giving us a combined twenty unique error profiles for our program. Therefore, with this process, no program will have the same errors from testing iteration to testing iteration, allowing for a random error removal process as errors are discovered.

10 Data Collection Using Elevator Controller

Previously in [2] we illustrated a four state testing procedure. To re-cap, this process occurs as follows: (i) Compilation, (ii) Static Analysis, (iii) Compilation After Static Analysis (CASA), and (iv) Testing. From each of these stages, errors are determined and removed, and the time taken to locate all errors is recorded for each stage. Each stage's time and error count is combined, respectively, giving us total errors found during a testing iteration. Testing iterations are repeated until no more errors are located in all four stages.

A database is utilized to store the errors found. We built our database in Microsoft Access as it comes configured to allow the use of style sheets for inputting data entries, and it was easy to build a single repository to include all runs for each error rate for each program. By incorporating slots for time, location in the code, and the categorization method from Putz [69], we have created a comprehensive form that shows the discovery and removal of each error in the system being tested. Figure 17 shows the entry form for the database, and Figure 18 shows an excerpt from the table of entries after errors have been entered.

This digital recording method is faster than trying to record errors by hand and allows the developer to focus on finding and removing errors in a timely manner. This helps to reduce bias in the data and exclude as much as possible extra time that might be taken up by writing the findings down or using other recording methods. This also allows the developer to have a digital repository of their error findings, allowing them to see possible improvement from program to program as they utilize our testing procedure.

11 Results and Data Analysis

Using the testing procedure described in Section 10 and explained in [2], we completed ten runs for each error rate on our starting example the elevator controller. The results [2, 3] showed that using our complete development process, from formal specification through testing, demonstrates that software systems can use error metrics that fit statistical models, as originally proposed by Musa [25, 26] and Drake [30]. Further refined models [29] were explored in [3] but were inconclusive as to whether they were better than the base exponential model. This required further testing.

Now we have completed the same testing procedure on our remaining benchmark programs, using feedback garnered from our work [3]. Critique of the work suggested improving the fit of the models. As the models are non-linear, traditional R-squared metrics do not apply, as these are just for linear regression models [78]. We did use the pseudo adjusted R-squared value along with the average error of regression, S, to determine the best fit for our models in [3]. To enhance these values while not degrading the accuracy of the results, we looked into mathematical methods for improvement. The best method seems to be the removal of outliers from the data set. Previously we had attempted the removal of outliers with the elevator controller; however this relied on what is considered the "eyeball test," looking for data that did not fit with the rest, and that is not robust. For our current approach at eliminating outliers, we focused on the recorded time to find all errors in the first testing iteration. To do this we take the total time for iteration one of all tests completed, ten values in total, and calculate the mean and standard deviation of these values. From these values we generate a range of one standard deviation centered around the mean. Any test run whose starting time falls plus or minus one standard deviation outside of the mean has the entire run purged from the results. In most of the benchmarks this results in one to two strings of data being removed, with no more than three being removed in the most extreme case. The test runs removed also vary in number of iterations completed to reach "zero" errors. For the Automated Door results, we can see from Table 7 that two data sets are removed from the results.

This process modifies the results without compromising the randomness of the data or the model. With this new subset of results, we calculate the exponential, modified exponential, and multinomial exponential models as done in [3]. Table 8 shows the S and pseudo adjusted R-squared for the base results and then the three model types after the data improvement.

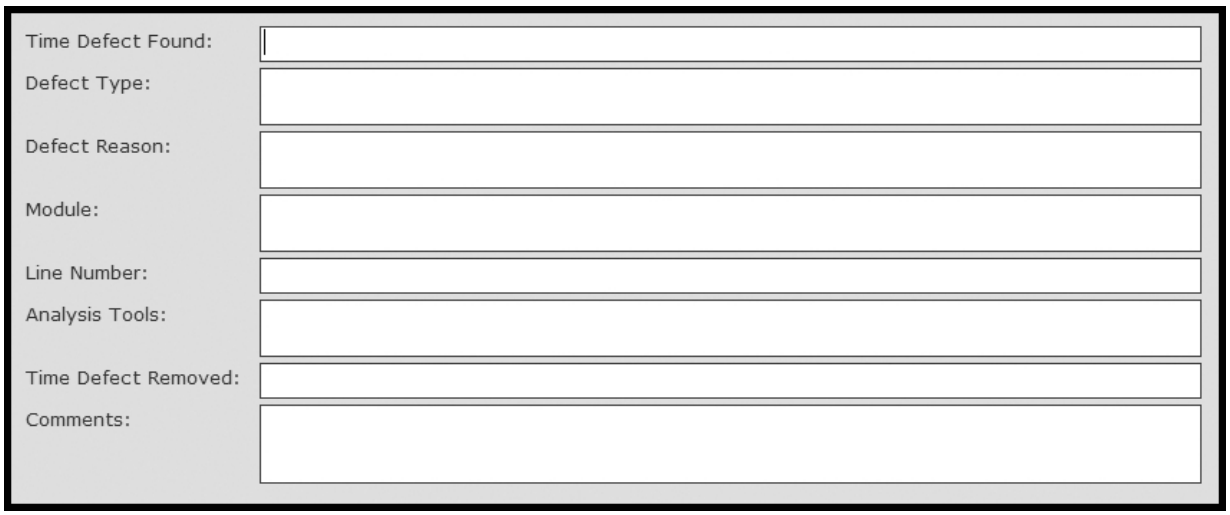


Figure 17: Form used to input errors into the database

Time Defect Found	Defect Type	Defect Reason	Module	Line Number	Analysis Tools	Time Defect Removed
12/3/2015 - 14:58:00.00						
12/3/2015 - 15:04	IU	KN	elevatorV1.	23	Compiler	12/3/2015 - 15:30
12/3/2015 - 15:05	WN	TY	elevatorV1.	25	Compiler	12/3/2015 - 15:38
12/3/2015 - 15:06	IU	OM	elevatorV1.	28	Compiler	12/3/2015 - 15:42
12/3/2015 - 15:07	IU	KN	elevatorV1.	31	Compiler	12/3/2015 - 16:02
12/3/2015 - 15:07	MI	OM	elevatorV1.	32	Compiler	12/3/2015 - 16:04
12/3/2015 - 15:08	WE	TY	eleavtorV1.	33	Compiler	12/3/2015 - 16:09
12/3/2015 - 15:09	WE	TY	elevatorV1.	45	Compiler	12/3/2015 - 16:09
12/3/2015 - 15:09	WN	TY	elevatorV1.	59	Compiler	12/3/2015 - 16:13
12/3/2015 - 15:10	WN	TY	elevatorV1.	70	Compiler	12/3/2015 - 16:14
12/3/2015 - 15:11	MI	TY	elevatorV1.	84	Compiler	12/3/2015 - 16:18
12/3/2015 - 15:12	WT	KN	elevatorV1.	90	Compiler	12/3/2015 - 17:08
12/3/2015 - 15:12	WE	OM	elevatorV1.	91	Compiler	12/3/2015 - 17:11
12/3/2015 - 15:13	WN	TY	elevatorV1.	111	Compiler	12/3/2015 - 17:13
12/3/2015 - 15:14	WE	OM	elevatorV1.	135	Compiler	12/3/2015 - 17:16
12/3/2015 - 15:14	WT	IG	elevatorV1.	145	Compiler	12/3/2015 - 17:19
12/3/2015 - 15:15	WN	TY	elevatorV1.	148	Compiler	12/3/2015 - 17:20
12/3/2015 - 15:15	WA	TY	elevatorV1.	149	Compiler	12/3/2015 - 17:22
12/3/2015 - 15:15	MI	OM	elevatorV1.	159	Compiler	12/3/2015 - 17:24
12/3/2015 - 15:16	WN	TY	elevatorV1.	160	Compiler	12/3/2015 - 17:28
12/3/2015 - 17:30	MI	OM	elevatorV1.	161	Compiler	12/3/2015 - 17:31
12/3/2015 - 17:33	WN	TY	elevator1V.	35	Compiler	12/3/2015 - 17:34

Figure 18: Excerpt from the table holding the errors discovered in run one of testing the elevator controller

Table 7: The automated door outlier removal process

First Iteration Starting Times		Mean of Starting Times	Standard Deviation of Starting Times	One Standard Deviation		Starting Times Without Outliers
47.005		27.525	8.35235	35.877		22.292
22.292				19.1727		24.296
24.296						28.060
28.606						28.223
28.223						28.846
28.846						34.313
34.313						19.502
19.502						24.636
24.636						
18.077						

Table 8: The final results for our benchmark programs

Error Rate	Program Tested	Data Model	S	Pseudo Adjusted R-Squared
High	Automated Door	Base Exponential	5.6586	0.4267
		Exponential	4.8513	0.5957
		Modified Exponential	4.7571	0.5789
		Multinomial Exponential	3.8516	0.7452
	Security System	Base Exponential	5.6457	0.6351
		Exponential	4.7965	0.7297
		Modified Exponential	4.7162	0.7201
		Multinomial Exponential	N/A	N/A
	Streetlight	Base Exponential	4.4259	0.5411
		Exponential	3.7957	0.6425
		Modified Exponential	3.9823	0.6356
		Multinomial Exponential	3.8650	0.6293
	Tollbooth	Base Exponential	5.2394	0.4518
		Exponential	4.9228	0.5655
		Modified Exponential	4.8191	0.5316
		Multinomial Exponential	4.5043	0.6149
	Vending Machine	Base Exponential	2.4275	0.8491
		Exponential	1.7077	0.9306
		Modified Exponential	1.8919	0.9248
		Multinomial Exponential	1.7077	0.9306
Low	Automated Door	Base Exponential	4.5366	0.4411
		Exponential	3.8664	0.6091
		Modified Exponential	4.1704	0.5866
		Multinomial Exponential	3.9120	0.6091
	Security System	Base Exponential	2.6590	0.8513
		Exponential	1.5579	0.9488
		Modified Exponential	1.5175	0.9457
		Multinomial Exponential	N/A	N/A
	Streetlight	Base Exponential	2.6988	0.7028
		Exponential	2.4116	0.7594
		Modified Exponential	2.3674	0.7488
		Multinomial Exponential	2.4524	0.7512
	Tollbooth	Base Exponential	3.0550	0.6349
		Exponential	1.9577	0.8601
		Modified Exponential	2.1923	0.8496
		Multinomial Exponential	1.1234	0.9539
	Vending Machine	Base Exponential	2.4878	0.6034
		Exponential	1.2017	0.8980
		Modified Exponential	1.3557	0.8902
		Multinomial Exponential	0.8980	1.2017

As we can see from the results, there is a definitive increase in the model fitment compared to the base, unmodified exponential model when the outliers are removed. The results show that in some cases the standard exponential model fits the data the best, in others the modified exponential, and in others the multinomial exponential. For the Vending Machine results, in both the high and low error rate results, the multinomial is the base exponential, as the best exponent has only the "x" term in it. From the results we can see that there is not any way to predict at this time which model works best for which rate, as each group has examples where each is the best for that program. In the case of the Security System, the multinomial could not be processed for both error rates, as the multinomial equation that fit best diverged to infinity when used as the exponent to an exponential.

12 Conclusion

The work presented here shows our complete design and testing process used in creating an initial benchmark suite for software reliability. Error analysis discovery with formal methods was shown to be more widely applicable than originally reported in [1], as we showed that errors could also be found by ATPs in the proof, not just the specification. We illustrated our RNG testing and random error generation that previously was mentioned in [2] but was not demonstrated in that work. This was utilized to help enhance our results in [3] and expand our tested programs. This expanded program suite illustrates that software errors, using a proper development process, can be shown to be modeled statistically similar to hardware. We have shown through the removal of outliers that the results can be improved further still. This develop-

ment process and the full results set a benchmark, so other development processes can be researched and even better software analysis can be achieved.

13 Future Work

Based on the work presented here, the first major area of expansion is the benchmarks themselves. We have shown the development and testing process allows small scale software programs to meet the statistical models that were originally theorized. This is great for embedded systems, but the process may be applicable to large system on chip solutions, or even large scale, big data applications. To begin this expansion, we will develop larger, more complex versions of the six benchmark programs already developed to see if the results hold when scaled up.

Similar to this area is the need to improve the development process with good, well developed software modules. In hardware, most systems are incorporated with ICs that come from manufacturers with known reliability. The same should be true of software. If modules have already been vetted through the development process, then they should be correct and accurate, allowing their reuse as a black box, and improving the error results of a new system prior to testing. This would be similar to pulling an IC of AND or NOT gates from the parts bin, knowing what it does and expecting it to work, without seeing the intricacies inside the black box.

The other side of safety critical systems is determining trust. Work in this space has looked into ways of showing trust for both hardware and software. Our development process could aid in this area, being twofold with reliability, if measures of security and trust can be incorporated into the formal method specification. In the future we will use the original elevator specification and redesign it, incorporating new security measures, while maintaining our current reliability measures. We will attempt to show the proofs can be developed to show security and reliability in one step, allowing for their incorporation from the beginning of design.

One final area of expansion is the integration of the software designed with this process into a targeted application. This requires further improvement of the software reliability in the system. Unlike hardware, which can use redundancy of circuitry to improve reliability, copying software with an undiscovered bug just propagates that bug to all versions. One solution may be to develop a hardware based monitor, possibly with machine learning, to determine when a software anomaly has been seen, and correct it. AI could be trained on the state and inputs of the system to know what outputs should be provided, and if the wrong output is generated from a known state, then a safety protocol could be enacted for a trusted system, or just replaced with the correct value for a non-critical system.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgments The authors would like to thank L-3 Fuzing and Ordnance Systems. A portion of this work was funded in part by their generous grant.

References

[1] J. Lockhart, C. Purdy, and P. A. Wilsey, "The use of automated theorem proving for error analysis and removal in safety critical embedded system specifications," in *2017 IEEE National Aerospace and Electronics Conference (NAECON)*, June 2017, pp. 358–361.

[2] J. Lockhart, C. Purdy, and P. A. Wilsey, "Error analysis and reliability metrics for software in safety critical systems," in *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Oct 2016, pp. 1–4.

[3] J. Lockhart, C. Purdy, and P. A. Wilsey, "Error analysis and reliability metrics for software in safety critical systems," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018.

[4] A. Mishchenko, S. Chatterjee, and R. Brayton, "Dag-aware aig rewriting: a fresh look at combinational logic synthesis," in *2006 43rd ACM/IEEE Design Automation Conference*, July 2006, pp. 532–535.

[5] C. Albrecht, "Iwls 2005 benchmarks," "", Tech. Rep., Jun 2005.

[6] J. Cong and K. Minkovich, "Optimality study of logic synthesis for lut-based fpgas," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 230–239, Feb 2007.

[7] Static & Formal Verification. [Online]. Available: https://www.synopsys.com/verification/static-and-formal-verification.html (Accessed 2017-3-17).

[8] A. J. Hu. Formal hardware verification with bdds: An introduction. [Online]. Available: https://www.cs.ox.ac.uk/files/4309/97H1.pdf (Accessed 2017-4-22).

[9] A. J. Hu, "Formal hardware verification with bdds: an introduction," in *1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PACRIM. 10 Years Networking the Pacific Rim, 1987-1997*, vol. 2, Aug 1997, pp. 677–682 vol.2.

[10] Arvind, N. Dave, and M. Katelman, "Getting formal verification into design flow," in *Proceedings of the 15th International Symposium on Formal Methods*, ser. FM '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 12–32. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68237-0_2

[11] Arvind, N. Dave, and M. Katelman. Getting Formal Verification into Design Flow. [Online]. Available: http://people.csail.mit.edu/ndave/Research/fm_2008.pdf (Accessed 2017-4-22).

[12] J. Markoff. (2016) Moore's law running out of room, tech looks for a successor. [Online]. Available: http://tinyurl.com/MooresLawOutOfRoom (Accessed 2016-5-6).

[13] J. Hruska. (2016, July) Moore's law scaling dead by 2021, to be replaced by 3d integration. [Online]. Available: https://tinyurl.com/EXT-MooresLaw (Accessed 2018-3-17).

[14] R. Smith. (2015) Amd dives deep on high bandwith memory - what will hmb bring amd. [Online]. Available: http://www.anandtech.com/show/9266/amd-hbm-deep-dive (Accessed 2016-4-29).

[15] AMD. (2015) High bandwidth memory — reinventing memory technology. [Online]. Available: http://www.amd.com/en-us/innovations/software-technologies/hbm (Accessed 2016-4-29).

[16] J. Hruska. (2017, July) Mit develops 3d chip that integrates cpu, memory. [Online]. Available: https://tinyurl.com/EXT-MIT-3D-Chip (Accessed 2018-3-17).

[17] H. Knight. (2017, July) New 3-d chip combines computing and data storage. [Online]. Available: http://news.mit.edu/2017/new-3-d-chip-combines-computing-and-data-storage-0705 (Accessed 2018-3-17).

[18] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-inspired stochastic computing using binary cbram synapses," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2402–2409, July 2013.

[19] K. Alrawashdeh and C. Purdy, "Fast hardware assisted online learning using unsupervised deep learning structure for anomaly detection," in *International Conference on Information and Computer Technologies (ICICT 2018)*. IEEE, 2018, p. In Press.

[20] T. Griggs and D. Wakabayashi. (2018, March) How a self-driving uber killed a pedestrian in arizona. [Online]. Available: https://tinyurl.com/ybvcgfu4 (Accessed 2018-3-28).

[21] R. Marosi. (2010) Runaway prius driver: 'i was laying on the brakes but it wasn't slowing down'. [Online]. Available: http://articles.latimes.com/2010/mar/10/business/la-fi-toyota-prius10-2010mar10 (Accessed 2014-1-26).

[22] A. Haxthausen, "An introduction to formal methods for the development of saftey-critical applications," Technical Univeristy of Denmark, Lyngby, Denmark, Tech. Rep., August 2010.

[23] G. Finzer. (2014) How many defects are too many? [Online]. Available: http://labs.sogeti.com/how-many-defects-are-too-many/ (Accessed 2016-4-29).

[24] V. R. Basili and B. T. Perricone, "Software errors and complexity: An empirical investigation," *Commun. ACM*, vol. 27, no. 1, pp. 42–52, Jan. 1984. [Online]. Available: http://doi.acm.org/10.1145/69605.2085

[25] J. D. Musa, "A theory of software reliability and its application," *IEEE Transactions on Software Engineering*, vol. SE-1, no. 3, pp. 312–327, Sept 1975.

[26] J. D. Musa and A. F. Ackerman, "Quantifying software validation: when to stop testing?" *IEEE Software*, vol. 6, no. 3, pp. 19–27, May 1989.

[27] S. Yamada, *Software reliability modeling: fundamentals and applications*. Springer, 2014. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-4-431-54565-1.pdf

[28] S. Yamada and S. Osaki, "Reliability growth models for hardware and software systems based on nonhomogeneous poisson processes: A survey," *Microelectronics Reliability*, vol. 23, no. 1, pp. 91 – 112, 1983. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0026271483913720

[29] S. Yamada, *Stochastic Models in Reliability and Maintenance*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, ch. Software Reliability Models, pp. 253–280. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24808-8_10

[30] H. D. Drake and D. E. Wolting, "Reliability theory applied to software testing," *Hewlett-Packard Journal*, vol. 38, pp. 35–39, 1987.

[31] RTCA and EUROCAE, "Software considerations in airborne systems and equipment certification," RTCA, Inc., Washington, D.C., USA, Tech. Rep., 1992.

[32] Federal Aviation Administration, "Advisory circular 20-174," U.S. Department of Transportation, Tech. Rep., 2011. [Online]. Available: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_20-174.pdf

[33] J. Lockhart, C. Purdy, and P. Wilsey, "Formal methods for safety critical system specification," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2014, pp. 201–204.

[34] M. Popovic, V. Kovacevic, and I. Velikic, "A formal software verification concept based on automated theorem proving and reverse engineering," in *Engineering of Computer-Based Systems, 2002. Proceedings. Ninth Annual IEEE International Conference and Workshop on the*, 2002, pp. 59–66.

[35] J. Rushby, "Harnessing disruptive innovation in formal verification," in *Software Engineering and Formal Methods, 2006. SEFM 2006. Fourth IEEE International Conference on*, Sept 2006, pp. 21–30.

[36] W. Visser, K. Havelund, G. Brat, and S. Park, "Model checking programs," in *Automated Software Engineering, 2000. Proceedings ASE 2000. The Fifteenth IEEE International Conference on*, 2000, pp. 3–11.

[37] Lemma 1, "Getting started," in *ProofPower: Tutorial*. Berkshire, United Kingdom: Lemma 1 Ltd, 2006, ch. 1, pp. 5–10.

[38] H. Ganzinger and V. Sofronie-Stokkermans, "Chaining techniques for automated theorem proving in many-valued logics," in *Multiple-Valued Logic, 2000. (ISMVL 2000) Proceedings. 30th IEEE International Symposium on*, 2000, pp. 337–344.

[39] C. Morgan, "Methods for automated theorem proving in non-classical logics," *Computers, IEEE Transactions on*, vol. C-25, no. 8, pp. 852–862, Aug 1976.

[40] A. A. Larionov, E. A. Cherkashin, and A. V. Davydov, "Theorem proving software, based on method of positively-constructed formulae," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 965–968.

[41] M. Kaufmann and J. S. Moore, "Some key research problems in automated theorem proving for hardware and software verification," in *RACSAM. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, vol. 98, no. 1, Oct 2004, pp. 181–195.

[42] J. Bowen and V. Stavridou, "Safety-critical systems, formal methods and standards," *Software Engineering Journal*, vol. 8, no. 4, pp. 189–209, Jul 1993.

[43] E. M. Clarke and J. M. Wing, "Formal methods: State of the art and future directions," *ACM Comput. Surv.*, vol. 28, no. 4, pp. 626–643, Dec 1996.

[44] Z. Li, S. Lu, S. Myagmar, and Y. Zhou, "Cp-miner: finding copy-paste and related bugs in large-scale software code," *IEEE Transactions on Software Engineering*, vol. 32, no. 3, pp. 176–192, March 2006.

[45] S. Morasca, "A probability-based approach for measuring external attributes of software artifacts," in *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, Oct 2009, pp. 44–55.

[46] A. Varshney, R. Majumdar, C. Choudhary, and A. Srivastava, "Role of parameter estimation & prediction during development of software using srgm," in *Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on*, Sept 2015, pp. 1–6.

[47] D. Tang and M. Hecht, "Evaluation of software dependability based on stability test data," in *Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers., Twenty-Fifth International Symposium on*, June 1995, pp. 434–443.

[48] Y. Tamura and S. Yamada, "Reliability analysis based on a jump diffusion model with two wiener processes for cloud computing with big data," *Entropy*, vol. 17, no. 7, pp. 4533–4546, 2015. [Online]. Available: http://www.mdpi.com/1099-4300/17/7/4533

[49] S. Yamada and M. Yamaguchi, "A method of statistical process control for successful open source software projects and its application to determining the development period," *International Journal of Reliability, Quality and Safety Engineering*, vol. 23, no. 05, p. 1650018, 2016. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218539316500182

[50] ISO/IEC, "Information technology: Z formal specification notation : syntax, type system, and semantics," International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland, Standard 13568, 2002.

[51] ISO/IEC, "Information technology, programming languages, their environments and system software interfaces, vienna development method, specification language," International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland, Standard 13817-1, 1996.

[52] M. Huth and M. Ryan, "Propositional logic," in *Logic in Computer Science: Modelling and Reasoning about Systems*, 1st ed. Cambridge, United Kingdom: Cambridge University Press, 2000, ch. 1, pp. 1–88.

[53] J. Davies and J. Woodcock, "Propositional logic," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 2, pp. 9–26.

[54] M. Huth and M. Ryan, "Predicate logic," in *Logic in Computer Science: Modelling and Reasoning about Systems*, 1st ed. Cambridge, United Kingdom: Cambridge University Press, 2000, ch. 2, pp. 90–146.

[55] J. Davies and J. Woodcock, "Predicate logic," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 3, pp. 37–2012 44.

[56] J. M. Spivey, "Basic concepts," in *Understanding Z: A Specification Language and its Formal Semantics*, 1st ed. New York, New York: Cambridge University Press, 1988, ch. 2, pp. 18–24.

[57] J. Davies and J. Woodcock, "Sets," in *Using Z: Specification, Refinement, and Proof*, 1st ed. Hertfordshire, United Kingdom: Prentice Hall Europe, 1996, ch. 5, pp. 57–72.

[58] J. M. Spivey, "Studies in z style," in *Understanding Z: A Specification Language and its Formal Semantics*, 1st ed. New York, New York: Cambridge University Press, 1988, ch. 5, pp. 98–113.

[59] Lemma 1, "Solutions to exercises," in *ProofPower: Z Tutorial*. Berkshire, United Kingdom: Lemma 1 Ltd, 2006, ch. 8, pp. 143–149.

[60] H. Pishro-Nik. ("", "") 11.1.4 nonhomogeneous poisson processes. [Online]. Available: https://tinyurl.com/yb5kbspn (Accessed 2018-12-7).

[61] K. Siegrist. ("", "") 6. non-homogeneous poisson processes. [Online]. Available: https://tinyurl.com/yc9y9vch (Accessed 2018-12-7).

[62] MIT OpenCourseWare. (2011, Spring) Chapter 2 poisson processes. [Online]. Available: https://tinyurl.com/ycg6r3n4 (Accessed 2018-12-7).

[63] NIST. ("", "") 8.1.7.2. non-homogeneous poisson process (nhpp) - power law. [Online]. Available: https://tinyurl.com/yclorwqd (Accessed 2018-12-7).

[64] S. Yamada, "Software reliability models," in *Stochastic Models in Reliability and Maintenance*, S. Osaki, Ed. Springer Berlin Heidelberg, 2002, pp. 253–280.

[65] V. Putz. (2000) The personal software process: an independent study. [Online]. Available: http://www.nyx.net/~vputz/psp_index/book1.html (Accessed 2013-10-30).

[66] H. Watts, "The personal software process (psp)," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU/SEI-2000-TR-022, 2000. [Online]. Available: http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=5283

[67] V. Putz. (2000) Program 1a. [Online]. Available: http://www.nyx.net/~vputz/psp_index/x341.html (Accessed 2017-02-16).

[68] V. Putz. (2000) Program 2a: Simple loc counter. [Online]. Available: http://www.nyx.net/~vputz/psp_index/x1547.html (Accessed 2017-01-26).

[69] V. Putz. (2000) Chapter 2. lesson 2: Planning and measurement. [Online]. Available: http://www.nyx.net/~vputz/psp_index/c1028.html (Accessed 2013-10-30).

[70] D. E. Knuth, *The Art of Computer Programming, Vol. 1*, 3rd ed. Addison-Wesley, 1997.

[71] D. E. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd ed. Addison-Wesley, 1998.

[72] Marsaglia, "Note on a proposed test for random number generators," *IEEE Transactions on Computers*, vol. C-34, no. 8, pp. 756–758, Aug 1985.

[73] P. L'Ecuyer and R. Simard, "Testu01 a software library in ansi c for emperical testing of random number generators," Département d'Informatique et de Recherche Opérationnelle Université de Montréal, Tech. Rep., 2013. [Online]. Available: http://simul.iro.umontreal.ca/testu01/guideshorttestu01.pdf

[74] P. L'Ecuyer and R. Simard, "Testu01: A c library for empirical testing of random number generators," *ACM Trans. Math. Softw.*, vol. 33, no. 4, Aug. 2007. [Online]. Available: http://doi.acm.org/10.1145/1268776.1268777

[75] M. E. O'Neill, "Pcg: A family of simple fast space-efficient statistically good algorithms for random number generation," Harvey Mudd College, Claremont, CA, Tech. Rep. HMC-CS-2014-0905, Sep. 2014.

[76] F. Neugebauer, I. Polian, and J. P. Hayes, "Building a better random number generator for stochastic computing," in *2017 Euromicro Conference on Digital System Design (DSD)*, Aug 2017, pp. 1–8.

[77] J. Lockhart, K. A. Rawashdeh, and C. Purdy, "Verification of random number generators for embedded machine learning," in *2018 IEEE National Aerospace and Electronics Conference (NAECON)*, In Press 2018.

[78] UCLA: Institute for Digital Research and Education. (2011) Faq: What are pseudo r-squareds? [Online]. Available: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/ (Accessed 2018-6-9).

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 247-251 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

An Improved Cross-Connection Abatement Algorithm with RSSI Using In-Band Magnetic Field Control in Densely Located LC Wireless Charger Environments

Nam Yoon Kim[1], Jinsung Cho[2], and Chang-Woo Kim[3,*]

[1]*Dept. of Marine Technology and Commissioned Officer, Dae Duk University, 34111, Republic of Korea*

[2]*Dept. of Computer Science and Engineering, Kyung Hee University, 17104, Republic of Korea*

[3]*Dept. of Electronic Engineering, Kyung Hee University, 17104, Republic of Korea*

ARTICLE INFO

Article history:
Received: 08 August, 2018
Accepted: 29 October, 2018
Online: 10 November, 2018

Keywords:
Wireless Power Transfer
Cross-connection error
Magnetic field control

ABSTRACT

In densely located loosely coupled wireless charger environments, cross-connection errors can occur when wireless chargers operate at the same time within the same wireless communication range. In this work, an effective algorithm is proposed to prevent cross-connection error. The algorithm based on the cross-connection abatement technique with the received signal strength indicator of out-band BLE communication has improved using an in-band magnetic field signal controlled by a pulse-width-modulation-like waveform. The experimental results obtained from a proper test set in a RF shield room verify that the wireless charging system using the proposed algorithm provides wireless charging services without cross-connection errors.

1. Introduction

At present, loosely-coupled (LC) wireless chargers are employed to support the simultaneous charging of a multiplicity of devices, such as laptops, tablets, and mobile phones in multiple wireless charging pads (power transmitting units: PTUs) and charging devices (power receiving units: PRUs) [1-3]. In densely located LC wireless charger environments, cross-connection errors can occur when PTUs and PRUs operate simultaneously in the same wireless communication range. The cross-connection errors are a kind of the control error between the main PTUs and neighboring PRUs due to the interconnection of the PTU/PRU communication. Figure 1 shows the concept of the cross-connection errors. When two (or more) PTUs with their own PRU are disposed closely together and operate at the same time, a control error may occur as one PTU's communication connects to the other PRU. If the PTU recognizes the battery information of the other PRU, over-power or low-power transmission can occur which may damage the PTU or prevent charging. In other words, the cross-connection error may cause the over-power or low-power transfer that may damage the PTU and PRU. In [4], we proposed an effective algorithm that focuses on the measured value of the received signal strength indicator (RSSI) to avoid cross-connection error. The RSSI, however, is heavily affected by nearby objects and by environments regardless of distance [5]; *i.e.*, an RSSI signal tends to fluctuate due to many external factors, such as obstacles, multipath fading, interference diffraction, and absorption. Moreover, when the distances between PTUs are small, it is difficult for a PTU to distinguish its own PRU from the others via only comparison of the RSSI values because the RSSI values become similar. Therefore, the algorithm using only the RSSI signals is not able to prevent communication connection errors in very dense environments.

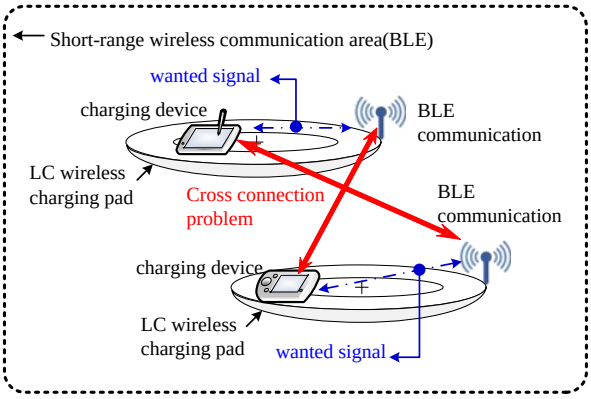


Figure 1. Conceptual diagram of the cross-connection errors between the charging devices and the wireless charging pads within the short-range wireless communication area.

*Chang-Woo Kim, 1732 Deogyoung-daero, Giheung-gu, Yongin, Gyeonggi-do, Republic of Korea, cwkim@khu.ac.kr

In this work, we propose an effective algorithm to solve the RSSI problems mentioned above. The proposed cross-connection abatement (CCA) algorithm is based on the CCA technique with the RSSI of out-band (2.4-GHz BLE) communication improved using the controlled magnetic field signal of in-band (6.78 MHz) power transmission. The CCA technique uses the out-band RSSI and in-band magnetic field signals progressively to enable the wireless chargers (PTUs) to be used in all dense environment without cross-connection errors. The PTU first finds the PRUs that are close to each other using the RSSI, then finds its own PRUs using the magnetic field signal controlled by the output power of the power amplifier. The PTU uses a waveform-controlled magnetic field signal, such as a pulse width modulation (PWM) signal, as a second signal to sense its PRU. To validate the performance of the proposed CCA algorithm, a test set was built in a dense charger environments.

2. System Description and Experimental Results

A test set was fabricated in a dense, loosely coupled (LC) measuring environment in a radio frequency shielded chamber with a 6.38 (length) m × 5.48 (width) m × 3.75 (height) m dimension [4]. The test set was composed of 10 wireless charging pads, 10 smartphones, a 16-V power on-off control switch, and a personal computer. The wireless charging pad with an area of 20 cm × 30 cm act as a PTU and the smartphone with a built-in wireless charging module acts as a PRU. The wireless charging pad that operates at 6.78 MHz with a 2.4-GHz Bluetooth low energy used to establish between PTUs and PRUs communication. The graphical user interface (GUI) monitors Bluetooth low energy RSSI values of all PRUs.

Figure 2 shows block diagrams of a PTU and PRU system used in the test set in a dense wireless-charger environment (two chargers within a 10-cm distance) and their output voltage waveforms as applied to the proposed CCA alorithm. The PRU is composed of a rectifier, a DC/DC converter, and a monitering circuit block. The rectifier converts received AC power signal to DC voltage. The DC/DC converter converts a received low voltage to a high voltage and the monitering block checks the charging conditions. The PTU is composed of a power amplifier, a buck converter, a wireless power on/off control switch, and a microprocessor. The power amplifier produces a 6.78-MHz RF power and the buck converter controls the output power, which produces a PWM-like waveform. For the PTU and PRU communication, a 2.4-GHz BLE was used. In [6], we proposed an effective CCA technique. The working sequence of the proposed technique is as follows.

To recognize the presence of a PRU in the PTU charging area, the wake-up power signal is transmitted over a regular cycle to sense the load. Once the advertisement signal is received, before the charging device is registered, an RSSI threshold level is identified to ensure that the PRU exists in the charging area. The RSSI threshold is measured as the position of the PRU in the PTU's charging area changes. The RSSI threshold level can be selected to be 3 dB (for noise margin) less than the minimum value of the highest RSSI level (for example, -12 dBm in [4]) among the received RSSI of PRUs. Once the threshold level is determined, PRUs with a value lower than that threshold is considered to have the cross-connection error. When this happens, the PTU turns off the wireless power for 100ms and returns to the load detection phase for a normal reconnection. The PRU with the RSSI threshold level registers its ID to initiate the wireless charging service. Then, the PTU monitors the load detection and advertisement signal of the PRUs in order to detect PRUs added or removed in real time during the wireless charging process.

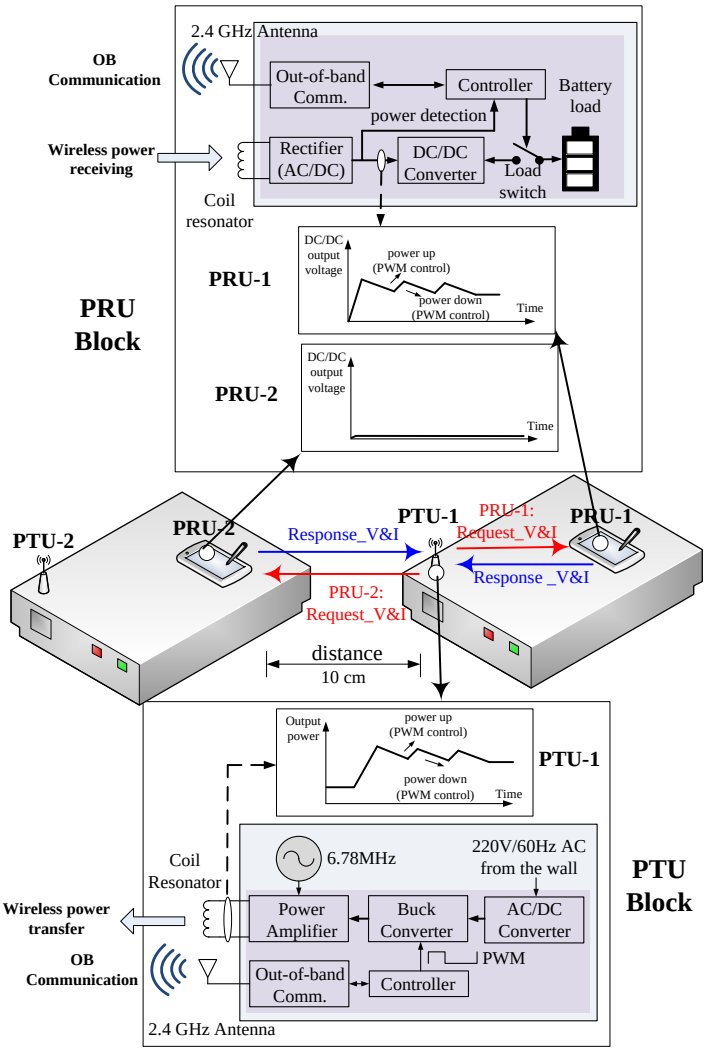
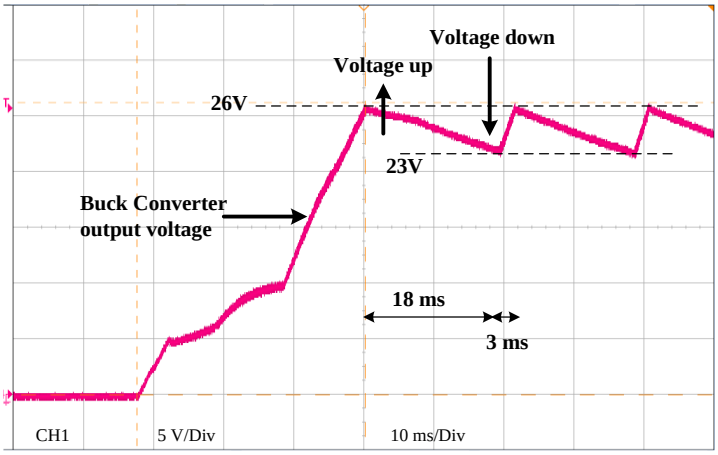


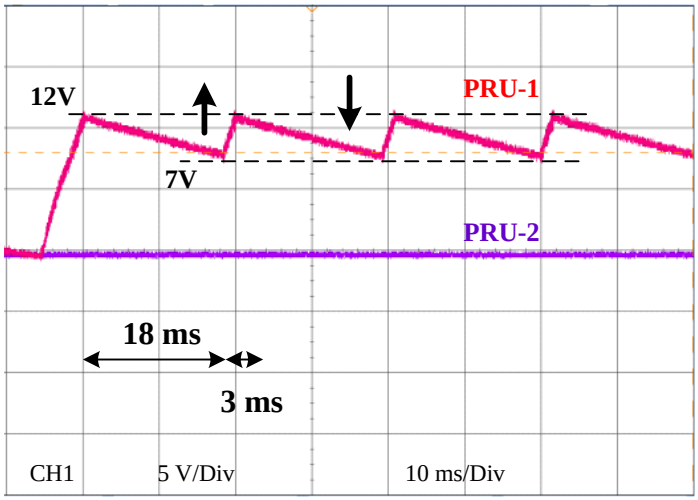
Figure 2. Block diagrams of a PTU and PRU sysytem used in the test set and schematic output voltage plots measured from PTU-1 and PRU-1, 2. PRU-1 belongs to the PTU-1, whereas PRU-2 is cross connected.

As the distance between two wireless charging pads (PTUs) decreases (less than 10 cm as shown in Figure 2), it becomes difficult for the PTU to distinguish its own PRU from the other one merely by comparing the RSSI values. This is because the received RSSI values of two PRUs eventually become similar to each other at the threshold level. For PRUs with RSSI values below the threshold, the communication connection is not made because the PTU identifies them on different charging pads. However, if the RSSI values of PRUs may be greater than or equal to the threshold, these PRUs should be revalidated for the cross-connection status. In that case, the PTU uses the magnetic field control of the transmit power signal to reaffirm the cross-connection status. A change in the magnetic field strength on the resonator of the PTU induced by varying the input current supplied from the power amplifier results in the change of the power information received by PRUs. The

PTU sends a magnetic field signal controlled by a waveform, such as a PWM-like waveform, which increases or decreases the output power of the power amplifier. This signal is then compared with the power change information received by the PRUs. If the power variation pattern is different from the PWM-like variation pattern, the cross-connection is confirmed.



(a)



(b)

Figure 3. Measured voltage waveforms of the PTU and the PRUs shown in Fig. 2. (a) Output voltage waveform of the power amplifier in PTU-1 (inset in Fig. 2). (b) Received voltage waveform at PRU-1 and PRU-2 (inset in Fig. 2).

Figure 3 shows the voltage waveforms measured at PTU-1, PRU-1, and PRU-2. Figure 3 (a) shows the output voltage measured at the Buck converter of PTU-1 which induced the magnetic field signal controlled by a waveform like PWM. Figure 3 (b) shows those at DC/DC converters of PRU-1 and PRU-2. In Figure 3 (a), the output voltage of the power amplifier varies from 23 V to 26 V with a 21-ms period. The rise time is 3 ms and the fall time is 18 ms. As shown in Figure 3 (b), the voltage change from 7V to 12V is repeatedly measured in PRU-1, which has a 21-ms period with a 3-ms rise time and 18-ms fall time. On the other hands the PRU-2 does not exhibit the voltage change. The measured results indicate that PRU-1 is a self-charging device of PTU-1, while PRU-2 belongs to another pad.

After PTU-1 identifies that PRU-2 is not its own PRU, PTU-1 disconnects communication with PRU-2 and connects to PRU-1 again to start the charging process. As a result, it is possible to accurately detect the occurrence of the cross-connection error by comparing the change in the intensity of the magnetic field of the PTU and the power change pattern received in the PRU.

Table 1 shows a comparison of the effectiveness of cross-connection abatement algorithms on the distance between PTUs. When the distance between PTUs is relatively long (longer than 15 cm), both techniques are very effective. As the distance is short (shorter than 10 cm), the CCA algorithm with RSSI and modulated magnetic field signal is still effective, while the CCA algorithm with only RSSI signal is not effective because the PTU does not recognize its own PRU.

Table 1. The comparison of effectiveness of cross-connection abatement algorithms on the distance between PTUs.

	Long distance between PTUs	Short distance between PTUs	Ref
CCA algorithm with only RSSI signal	effective	not effective	[4]
CCA algorithm with RSSI & modulated magnetic signals	effective	effective	This work

3. Improved Algorithm Description

Figure 4 shows the process flow of the proposed CCA algorithm with the RSSI and the controlled magnetic field signals. As shown in the figure, the left line (PTU) shows the command frames and the power levels (dark areas) of the PTU, and the center line (cross-connected PRU) and right line (own PRU) show the response and report frames, respectively, according to the request of the PTU. This process flow operates in four states: a load detection state, cross-connection prevention state, registration state, and charging state. The basic operation sequence of the CCA algorithm is as follows.

In the load detection state, the wake-up power supplies 5 V to the PA for 500 ms (t_{BEACON}) for every 3 sec (t_{BEACON_PERIOD}) cycle. The wake-up power is the minimum power level necessary to activate the micro control unit (MCU) of the PRUs to establish the communication. When a PRU is detected, the MCU of the PTU increases the voltage supply at the PA to 7 V to supply sufficient power for the PRU communication and the stable operation of the MCU. The woken-up PRUs proceed to the next stage, namely, the cross-connection prevention state. After the load detection between the PTU and PRUs, the PTU transmits a "request RSSI" frame to the PRUs in order to verify the state of cross connection. Upon receipt of the "request RSSI" frame, the woken-up PRUs immediately transmit their RSSI values to the PTU. By comparing the RSSI values with the threshold value (−15 dBm in [4]), the PTU first recognizes those PRUs as the possible PRUs that have greater RSSI values than the threshold value, before proceeding to accurately judge them as its own PRUs. In the case of PRUs that have lower RSSI values than the threshold value, no communication connection is made because the PTU recognizes them as belonging to other charging pads (PTUs). The PRUs that have higher RSSI values than the threshold value should be rechecked for the state of cross connection because RSSI values can become greater than the threshold value when the distance between the charging pads decreases. The PTU rechecks the state

of cross-connection using the magnetic field control of the transmitting power signal. The change in the magnetic field strength on a resonator of the PTU, induced by changing the input current supplied from a PA, changes the power (voltage and current, V&I) received by the PRUs. The power (V&I) change information received by the PRUs is reported through the "report V&I" frames whenever the "request V&I" frames are transmitted during this state.

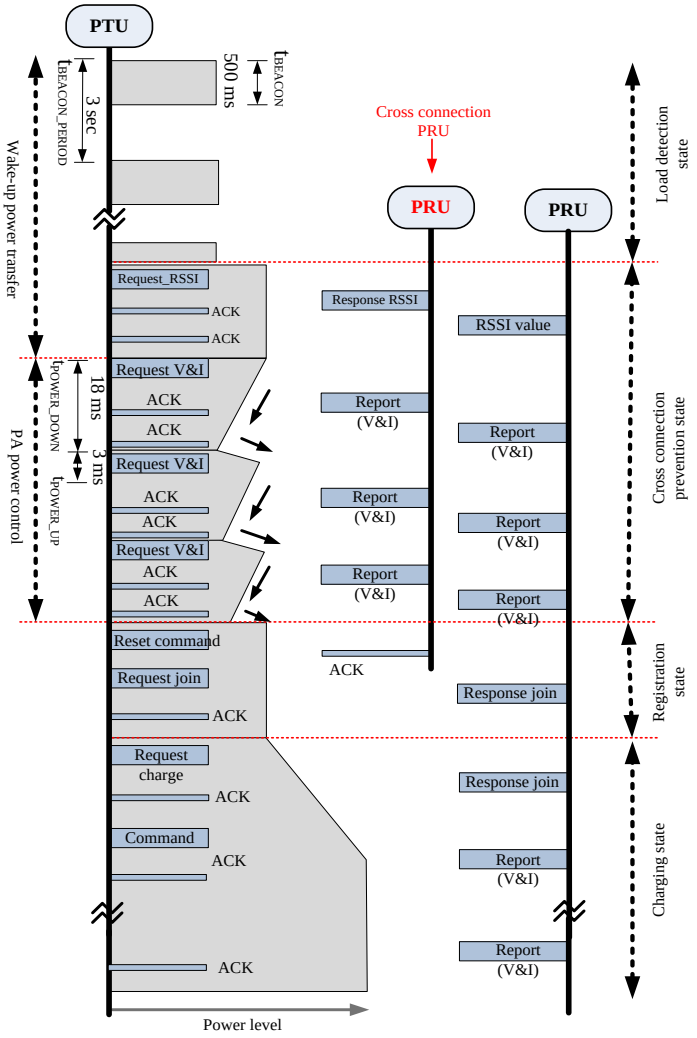


Figure 4. Sequence diagram of the proposed algorithm for cross-connection abatement.

The PTU causes an increase or decrease in the output power of the power amplifier and compares it with the power change information received by the PRUs. If these two power variation patterns are different, then a cross connection has occurred. In the algorithm, the magnetic field control is implemented for 18 ms (t_{POWER_UP}) for a power increase at the PA and for 3 ms (t_{POWER_UP}) for a power reduction at an interval of 21 ms. The control is carried out three times in a sequence. To detect the power change information (V&I) received by the PRUs based on the magnetic field control, the PTU transmits the relevant "request V&I" frames. Upon the receipt of the "request V&I" frames, the PRUs transmit the measured V&I information through the "report V&I" frames. The communication between the PTU and the PRUs occurs during the timespan of 21 ms of the power change. The three repetitions of the power-level change at the PTU and the power change information received by the PRUs are compared, and if the variation patterns turn out to be different or there are no changes in the received power, the PRUs start detecting places where cross-connection errors have occurred.

Consequently, the PTU transmits a reset command to the PRUs (such as the PRU-2) where cross-connections have occurred to disconnect the communication. The PRUs receive the reset command and switch to a standby state in order to receive the "request" frames from other PTUs. The PRUs (such as the PRU-1) that are verified to be its own proceed to the registration state.

After detecting the PRUs that are to be charged on its own pad in a cross-connection prevention state, the PTU sends a subscription request to its network and receives a session ID and the charging permit. The PTU then transmits the "request join" frames to its own PRUs. Consequently, the PRUs transmit the battery capacity information, which is necessary for charging, via the "response join" frames. By comparing the power capacity that the PTU can transmit and the battery capacity of the registered PRUs, the PTU decides whether the charging should be allowed. If the power transfer capacity exceeds, charging is not allowed, and the PRUs revert to the standby state. This information is notified to the PRUs through the acknowledgement (ACK) frames. After this registration state, the PTU proceeds to the charging state and starts a charging process. During the charging state, the PTU transmits the "charge start" commands to relevant PRUs to switch them to the charge state for charging the batteries. The PRUs switch on the load attached to the battery path and receive the power. The PTU monitors whether the power is delivered properly to the PRUs through the "report V&I" frames. Based on the power information received from PRUs, the PA adjusts its output power. The PTU sends back the "report" frames several times so that the power can be delivered stably to the PRUs. When the charging power is delivered to the PRUs, the voltage and current received by the battery are continuously monitored so that the required power can be controlled according to the charge mode. If the monitored voltage of the battery reaches a buffer voltage, the PTU judges that the charge has been completed. Thereupon, the PTU issues the "charge finish" commands to PRUs and switches the load connected to the battery to the off state, along with the transmission of an ACK command before displaying a buffer to the users through a user interface.

4. Conclusions

The proposed CCA algorithm has shown that cross-connection errors between wireless charging pads and charging devices in BLE wireless coverage can be prevented using simple RSSI threshold level detection and magnetic field control for power signal transmission. This approach can provide a reliable loosely coupled wireless charging service without an abrupt increasing in transmit power at the start of charging due to the gradual power-up pattern of the controlled magnetic field.

Acknowledgment

This work was supported by the Basic Science Research program through the National Research Foundation Korea funded by the Ministry of Education (NRF-2015R1D1A1A02061041) and by the Ministry of Science and ICT Korea under the ITRC (information technology research center) support program (IITP-2016-R2718-16-0012) supervised by the IITP (national IT industry promotion agency).

Reference

[1] J.J. Casanova, Z.N. Low, and J. Lin, "A loosely coupled planar wireless power system for multiple receivers," IEEE Trans. Ind. Electron., vol. 56, no. 8, pp. 3060-3068, Aug. 2009. https://doi.org/10.1109/ TIE.2009.2023633

[2] R. Tseng, B. von Novak, S. Shevde, and A.K. Grajski, "Introduction to the alliance for wireless power loosely coupled wireless power transfer system specification, Version 1.0," in 2013 IEEE WPT, May 2013, pp. 79−83. https://doi.org/10.1109/WPT.2013.6556887

[3] I. Yasar, L. Shi, K. Bai, X. Rong, Y. Liu, and X. Wang "Mobile phone mid-range wireless charger development via coupled magnetic resonance," in 2016 IEEE ITEC, June 2016, pp.1–8. https://doi.org/ 10.1109/ITEC.2016.7520217

[4] N.Y. Kim, S.-W. Yoon, and C.-W. Kim, "Cross-connection abatement in dense loosely coupled wireless charging pad environments," Electronics Lett., vol. 50, no. 6, pp. 461-462, Mar. 2014. https:// doi.org/10.1049/el.2013.3754

[5] B.-G. Lee and W.-Y. Chung, "Multi-target three-dimensional indoor navigation on a PDA in a wireless sensor network," IEEE Sensors Journal, vol. 11, no. 3, pp. 799–808 Mar. 2011. https:// doi.org/10.1109/JSEN.2010.2076802

[6] N.Y. Kim, J. Cho, and C.-W. Kim, "An effective technique for preventing cross-connection errors in dense loosely-coupled wireless charging pad environments," in 2017 ICUFN, July 2017, pp623-626. https://doi.org/10.1109/ICUFN.2017.7993866

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 248-257 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Observing and Forecasting the Trajectory of the Thrown Body with use of Genetic Programming

Konstantin Mironov[*,1,2], Ruslan Gayanov[2], Dmiriy Kurennov[1]

[1]*Institute of New Materials and Technologies, Ural Federal University, 620002, Russia*

[2]*Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, 450008, Russia*

ARTICLE INFO

Article history:
Received: 06 January, 2019
Accepted: 28 January, 2019
Online : 20 February, 2019

Keywords:
Robotic Catching
Genetic Programming
Forecasting
Machine Vision
Machine Learning

ABSTRACT

Robotic catching of thrown objects is one of the common robotic tasks, which is explored in a number of papers. This task includes subtask of tracking and forecasting the trajectory of the thrown object. Here we propose an algorithm for estimating future trajectory based on video signal from two cameras. Most of existing implementations use deterministic trajectory prediction and several are based on machine learning. We propose a combined forecasting algorithm where the deterministic motion model for each trajectory is generated via the genetic programming algorithm. Object trajectory is extracted from video sequence by the image processing algorithm, which include Canny edge detection, Random Sample Consensus circle recognition and stereo triangulation. After that rajectory is forecasted using proposed method. Numerical experiments with real trajectories of the thrown tennis ball show that the algorithm is able to forecast the trajectory accurately.

1. Introduction

This paper is an extension of the article presented at the IEEE International Symposium on Signal Processing and Information Technology [1]. Here and there we address the task of robotic catching of thrown objects or, more precisely the subtask of observing and predicting the trajectory of the thrown object.

With the development of robotics, mechanical systems acquire more and more features that were previously only available to humans. One of these possibilities is the ability to catch objects thrown in the air. Initially, a robotic capture of objects thrown in the air was described in 1991 in [2]. Later, this task was considered several times in a number of articles [1,3,11].

In addition to theoretical value, such a skill can have a practical application. For example, in light industry, the task of transporting workpieces between machine tools processing them often arises. The traditional solution to this problem is the use of various conveyor systems. Robotic transfer as a method of such transportation was proposed in 2006 by Frank [12]. This application was developed in [13-21]. Transportation of an object from some point of departure A to some destination B is as follows: the robot thrower located in A throws the object in direction B and notifies about it via the communication line, and the robot catcher located in B, having received the notification, carries out object capture on the fly.

The authors of [12,15] specify the following potential advantages of robotic throw compared to traditional conveyor-based systems:

- Greater flexibility. Flexibility is understood as the ability to quickly deploy, collapse and redevelop a transport network with an arbitrary topology, or to use it in production facilities with an arbitrary layout.
- Higher speed of object transportation.
- Reduced energy consumption.

The share of successful captures in most existing systems does not exceed 80% (two exceptions are described in [11] and [16]; in the first article, a high proportion of successful captures is provided by large linear dimensions of the gripping device; in the second, by throwing cylindrical objects of high aerodynamic stability), which is not sufficient for use in a real industrial environment. Thus, the practical implementation of transportation by robotic throw is a complex and relevant scientific task.

For a successful capture, it is necessary to know in which point of space the object will be at the moment of capture, and at what speed it will move at the same time [15]. The point in space and time where the capture is carried out is selected among the set of points that the object passes when it flies through the working space of the capture device. Their combination forms the trajectory of the object in the working space of the capture device. This

*Corresponding Author: Konstantin Mironov, mironovconst@gmail.com

trajectory must be predicted in advance so that the robot catcher has time to complete the capture [15]. Prediction is based on measuring the object trajectory immediately after the throw. In general, in [15], the following four subtasks are distinguished when ensuring the transportation of objects by robotic transfer:

- throw;
- capture;
- forecasting;
- tracking.

Here we consider last two subtasks. Trajectory forecasting is needed in order to provide the catcher with the information about object trajectory within the workspace of the gripper. Most of the trajectory forecasting algorithms are based on ballistic modeling of the flight. These models include the influence of gravitation only (e. g. [2,3,7]; in this case forecasting may be implemented by fitting a parabola to the reference of measured positions) or gravitation and air drag (e. g. [4,8,9]). This modeling requires preliminary knowledge about ballistic properties of thrown objects. However, human children do not need such a knowledge to catch the ball successfully. They do it only based on the previous experience. This circumstance motivated the development of learning-based forecasting algorithms, such as neural network trajectory predictor [18] and k nearest neighbor's trajectory predictor [19-21]. Learning-based techniques require collecting the sampling of trajectories in order to train the predictor. Here we propose the method, which lies between model-based and learning-based model. The predictor is using equations to define future positions of the object, but the parameters of these equations are obtained by the learning procedure of genetic programming. Learning does not require a sampling of past trajectories: the parameters of the model are learned from the initial part of the current trajectory.

We do not consider the task of providing correct throwing and catching movement in this article. This is a complex control task, solved by various works in the field of robotics and mechatronics, e.g. Implementation of robotic control within our project is discussed in [17].

A tennis ball is considered as the object to be thrown. On one hand, this object is quite complex and unstable aerodynamically [22] so that its trajectory cannot be accurately predicted using simple models; on the other hand, its aerodynamic characteristics are investigated in sufficient detail ([22] provides a detailed overview of its characteristics completed in 50 years) so that the aerodynamic model can be used to verify the accuracy of the algorithm functioning.

2. Extraction of the spatial coordinates from video signal

Tracking the trajectory of a moving object is a task that often arises in machine vision applications. Following examples could be mentioned: In our case, it is considered for the following conditions: the object is a sphere thrown at a speed of several meters per second at certain angle to the horizon. Such conditions are determined by the task of robotic capture of a thrown object in the system of transportation of objects by transfer.

Since monitoring is performed through a camera, tracking an object becomes the task of processing images and video. Positioning the flying ball in space is performed using stereo vision. The spatial position of a certain point is determined on the basis of its pixel coordinates on images from two cameras and on the basis of the system parameters: the relative location of the cameras, their focal lengths, etc. [23]. Camera parameters are configured using Zhang's calibration procedure [24-27].

The study of the stereo positioning accuracy is poorly described in the literature. Most of the articles describe positioning of static objects, for example, [28]. When positioning an object, errors inevitably occur. According to the classification proposed by Lee [28], they are divided into three types:

- Calibration errors. They are related to errors by calibration, i.e., in determining the parameters of the camera system. These errors are systematic and amount to no more than one millimeter per meter of distance.
- Quantization errors. They are associated with the transition from pixel to metric coordinates. The set of pixel coordinates does not correspond in space to a point, but to a certain area, the size of which increases with increasing distance from cameras. At a distance of up to two meters, the magnitude of quantization errors is small; at a greater distance, it becomes significant.
- Image processing errors. These errors are related to incorrect operation of image processing algorithms which are used to determine the position of a pixel point.

The influence of these errors on the positioning of a static spherical object was investigated in [26]. The object is tracked using two IDS uEye UI-3370CP [29] video cameras combined into a stereo pair. The resolution of each camera is 2048 by 2048 pixels. They are installed at a distance of several tens of centimeters. Studies of calibration and quantization errors showed that standard deviations due to calibration and quantization errors are less than 1.5 millimeters (ranges from 1 to 1.4 mm, with an increase in the range from 0.5 to 2.5 mm). Errors of image processing were more significant: the total standard deviation of the positioning of the sphere is up to 2.2 mm. Errors in the positioning of a thrown object in flight were analyzed in [27]. The algorithm described below was implemented in C ++ using the CUDA library, which allows deparallelizing of the calculations for their execution on the graphics processor [30]. Some extensions of the algorithm help improve positioning accuracy.

The procedure for determining the spatial coordinates of an object is illustrated in Figure 1 (the original images are shown in the first row). It includes the following steps:

- Background subtraction. The results are shown in the second row of Figure 1.
- Selecting boundaries using the Canny algorithm [31]. Border images are shown in the third row on Figure 1. .
- Circle detection on border images. The result of this stage are pixel coordinates of the circle center on each image. In [30], two methods of such prediction are compared: the Hough transformation [32,33] and the RANSAC [34] method. As a result, an algorithm based on the RANSAC method was chosen. It has the same accuracy as the Hough transformation, but requires fewer resources [30]. The algorithm selects three random points in the image, builds a circle on their basis and checks whether other points of the boundary image fit into this circle. This action is repeated sequentially until a circle is found that fits well

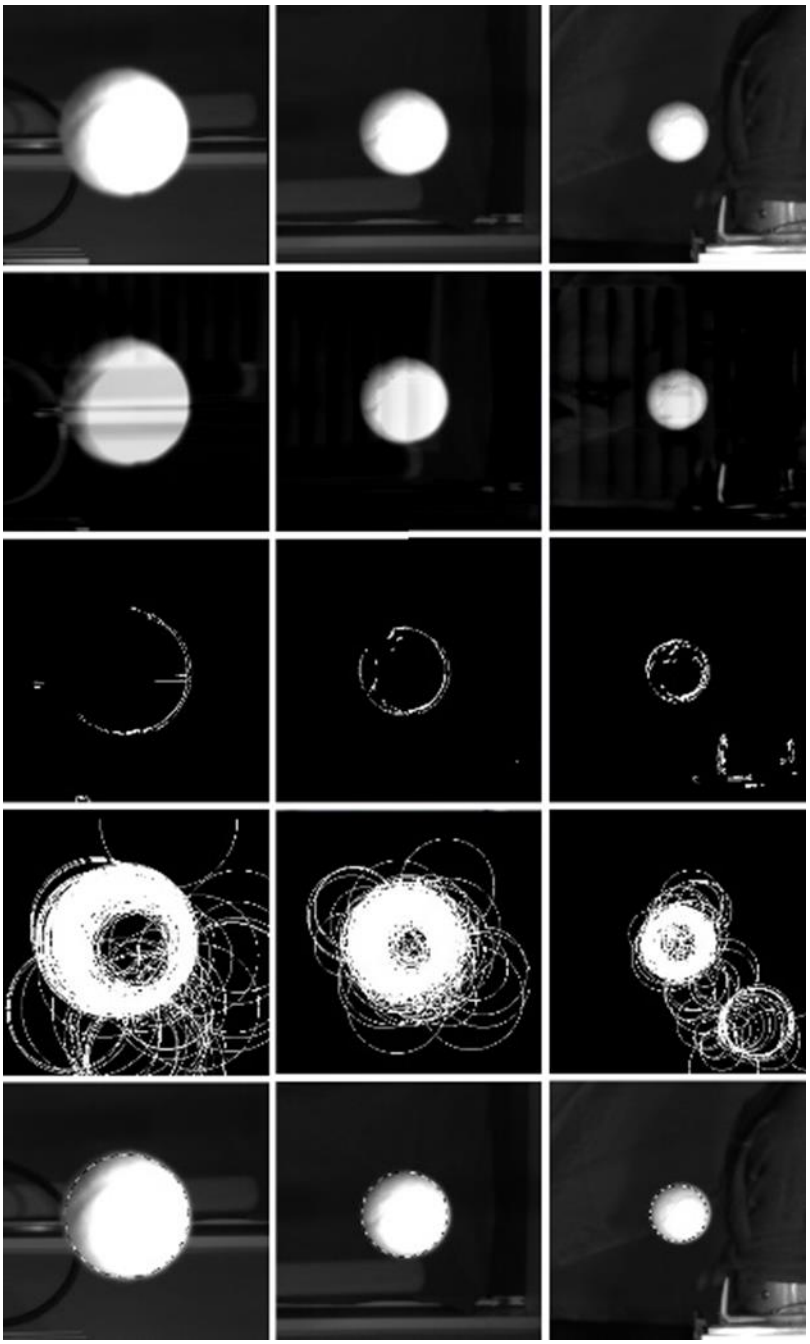


Figure 1. Circle recognition with RANSAC for three images.

with the points of the boundary image. The fourth row in Figure 1 shows the hypothetical circles generated by the algorithm, and the fifth row shows the selected circle projected onto the original image.

- Stereo triangulation. This is the operation of determining spatial coordinates for the center of an object based on its pixel coordinates in two images and using camera calibration parameters.

Coordinates obtained as a result of stereo-triangulation are then transferred to the system defined as follows:

- The center of coordinates coincides with the position of the object at the time of the throw.
- One of the axes is directed vertically upwards.
- The second axis is aligned with the horizontal projection of the direction of the throw.
- The transfer of coordinates into such a system provides a two main advantages. First, three-dimensional coordinates can be replaced by two-dimensional ones. Second, approximation of the trajectory by the plane allows you to identify outliers, i.e., filter out frames on which the position of the object is measured incorrectly. Image processing errors are associated with incorrect results in the first two steps, while calibration and quantization errors affect the result of stereo triangulation. Coordinate transform is described more precisely in the end of this section.

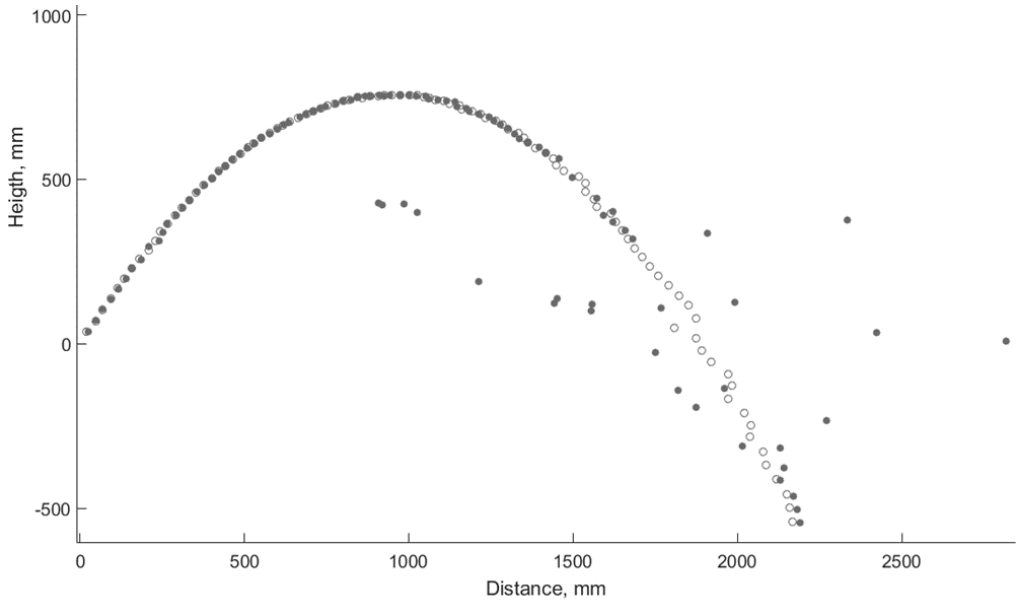


Figure 2. Graph of the trajectory measured with subtraction of the background (circles) and without it (ring)

The experiments described in [27] mainly concern the theoretical assessment of positioning accuracy. In [28], the author analyzed the errors found in the real situation. This analysis is complex, since there is no true data on the position of a real object at any given time. Many errors can be detected because they distort the smoothness of the path. This curve cannot be accurately determined analytically, but it is smooth [28]. Another way to assess errors is to approximate the measured values by a simplified model of the object movement. These models are not accurate, however, if they are more accurate than the vision system, the quality of the approximation can provide information about the accuracy of the observer.

Errors of individual processing steps can be detected by visual analysis of intermediate images. The quality of border selection can be assessed by comparing the found boundaries with the boundaries of the ball on the original image. The quality of circle recognition can be estimated by projecting the circles found on the original images. For example, a visual analysis of the images in Figure 1 shows that the border detection algorithm introduces some noise, but the RANSAC assessment gives plausible results. The disadvantage of this visual analysis is that it is performed by humans and cannot transmit objective information. However, it does detect some obvious tracking errors.

Subtracting the background before running the Kenny algorithm is an optional step, but in practice, this step is necessary for correct positioning of the object at a great distance. If background subtraction is not applied, the deviations of the measured values increase significantly when the distance from the camera to the object exceeds 1.5 meters. The effect is illustrated in Figure 2. Charts are shown for the same trajectory extracted by the RANSAC algorithm with and without background subtraction. You can see that at a distance of about 1.5 meters, the measurements almost coincide, and the trajectory looks like a second-order curve. Measurements with background subtraction (rings) retain this view afterwards, but measurements without subtracting the background (circles) become chaotic. This behavior is typical of most trajectories in a dataset.

Another way to obtain more accurate data for comparison and verification is typical for RANSAC. Since RANSAC does not provide the same results for different starts, several starts give several hypotheses about the position of the ball center. A correct statistical estimate based on these hypotheses is more accurate than the result of a single run of RANSAC. The results of multiple measurements are noisy and are not supported by the model of true motion and previous statistical knowledge, for example, the probability density function. According to [35], the least squares estimate shall be used under those conditions. Such an estimate for a static parameter with unknown random noise is equal to the average measurement result. In this paper, the mean value is replaced by the median. The median and average scores give similar results, but the median score is more resistant to emissions. The median of 1000 RANSAC launches was used in compiling the training base of trajectories; a further increase in the number of launches does not change the results of the median estimate. The use of such an estimate in real time is impossible due to the large amount of computation. An existing graphics processor can perform one run of RANSAC in real time (i.e., less than 9 ms for two images and less than 1 second for the entire trajectory). It takes about 10 minutes to run the RANSAC algorithm 1000 times.

The numerical evaluation of errors is given in Table 1. Here, the coordinates extracted by a single RANSAC run are compared with the results of the median estimate for 1000 runs. Differences are considered "errors." These numbers are not equal to real positioning errors, but they can be used to perceive the dispersion of measurements. Based on these differences, the standard deviation is calculated for each frame when the ball was thrown. In the table, each frames are combined into block to save space. Standard deviations are summarized based on 111 trajectories.

It can be seen that after the 65th frame, the parameter begins to increase strongly, and this growth is more impressive for the variant of the algorithm without subtracting the background. The reason for this increased stability at the beginning is that for the initial frames the size of the ball is larger and almost completely covers the image (compare the first and third columns in Figure 1). Therefore, the background borders make smaller distortions by the results of the border detection.

Table 1. Comparing the difference in millimeters between the measured 3D positions based on one RANSAC run and the median of 1000 RANSAC runs, for variants of the algorithm with and without background subtraction.

Frame Number	Standard Deviation		Median Error	
	Without Background Subtraction	With Background Subtraction	Without Background Subtraction	With Background Subtraction
1..5	7.9	6.4	1.6	0.8
6..10	4.0	1.9	1.8	0.9
11..15	3.7	2.1	2.0	1.1
16..20	2.8	1.9	1.6	0.5
21..25	2.1	1.8	1.4	0.3
26..30	4.0	2.2	2.2	0.5
31..35	22.1	17.2	2.9	2.2
36..40	10.9	4.0	3.4	0.4
41..45	24.8	3.2	3.7	0.4
46..50	29.8	3.9	4.5	0.7
61..65	63.9	14.8	5.5	1.0
66..70	187.4	41.3	10.5	4.2
71..75	305.6	138.5	20.4	5.4
76..80	520.4	242.2	208.2	7.3
81..85	897.0	229.3	171.9	8.0
86..90	1361.6	197.5	163.9	8.4
91..95	1450.0	212.1	176.1	9.3

It can be seen that even for the option with background subtraction, the standard deviation after the 70th frame reaches very high values. Standard deviation may not be the best option, as it has low emission resistance. Therefore, columns in the right-hand part of the table show median differences for the same blocks. The median results look the same as for standard deviations, but they are more detailed. For the algorithm without background subtraction, the average error lies at 3σ interval for static spheres, estimated as 6.75 mm [26], up to the 60th frame. For the algorithm with background subtraction, this property is preserved up to the 80th frame. In the version without background subtraction, an average value of more than 20 cm is reached after the 75th frame. This means that most frames are outliers in this area. Thus, measurements without background subtraction are practically useless.

Position measurements can be divided into inliers and outliers. Outliers are defined as measurements that are completely useless, even harmful for trajectory restoring. Inliers may be wrong, but they help to improve the score. Obviously, it is not possible to determine with 100% certainty whether a measurement is an inlier or outlier. The huge difference between the standard deviation and the median error at the end of the trajectory shows that outliers make up a large proportion of the measurements.

Trajectory construction demonstrates specific properties of these errors. Figure 3 shows three graphs describing the trajectory: relationship between the height of the object and the distance from the camera (upper graph), dependence of height from the frame number (bottom left) and the dependence of distance from the frame number (bottom right). It is easy to see that the first and third graphs appear to be noisy on the right side, and the height--time dependence retains the appearance of a smooth curve of second order. In other words, errors are mainly related to distance measurement.

The reason for error localization in one dimension is that when the distance to the object exceeds the distance between the cameras of the stereo system, one pixel has a greater influence on the measurement of distance from the camera than on the measurement of other coordinates. An illustration of this property is shown in Figure 4. Experiments show that the error value at large distances is significant for tracking in terms of measuring distance. This problem can be overcome by increasing the number of cameras used for tracking, but this can be costly.

The cameras must be located so that the effect of large distance errors on the quality of the system function is minimized. The following question should be answered. In which part of the path is accurate positioning the most important? During the first experiments, the cameras were located opposite the throwing device. In this situation, positioning in the first frames is the least accurate. It was possible to accurately position the ball from the 10th or 12th frame. In further experiments, the cameras were moved to the side of the throwing device. In this case, positioning in the initial part of the trajectory is quite accurate, but the final part of the trajectory is measured with higher errors. High accuracy in the initial part of the trajectory and lower accuracy in the final

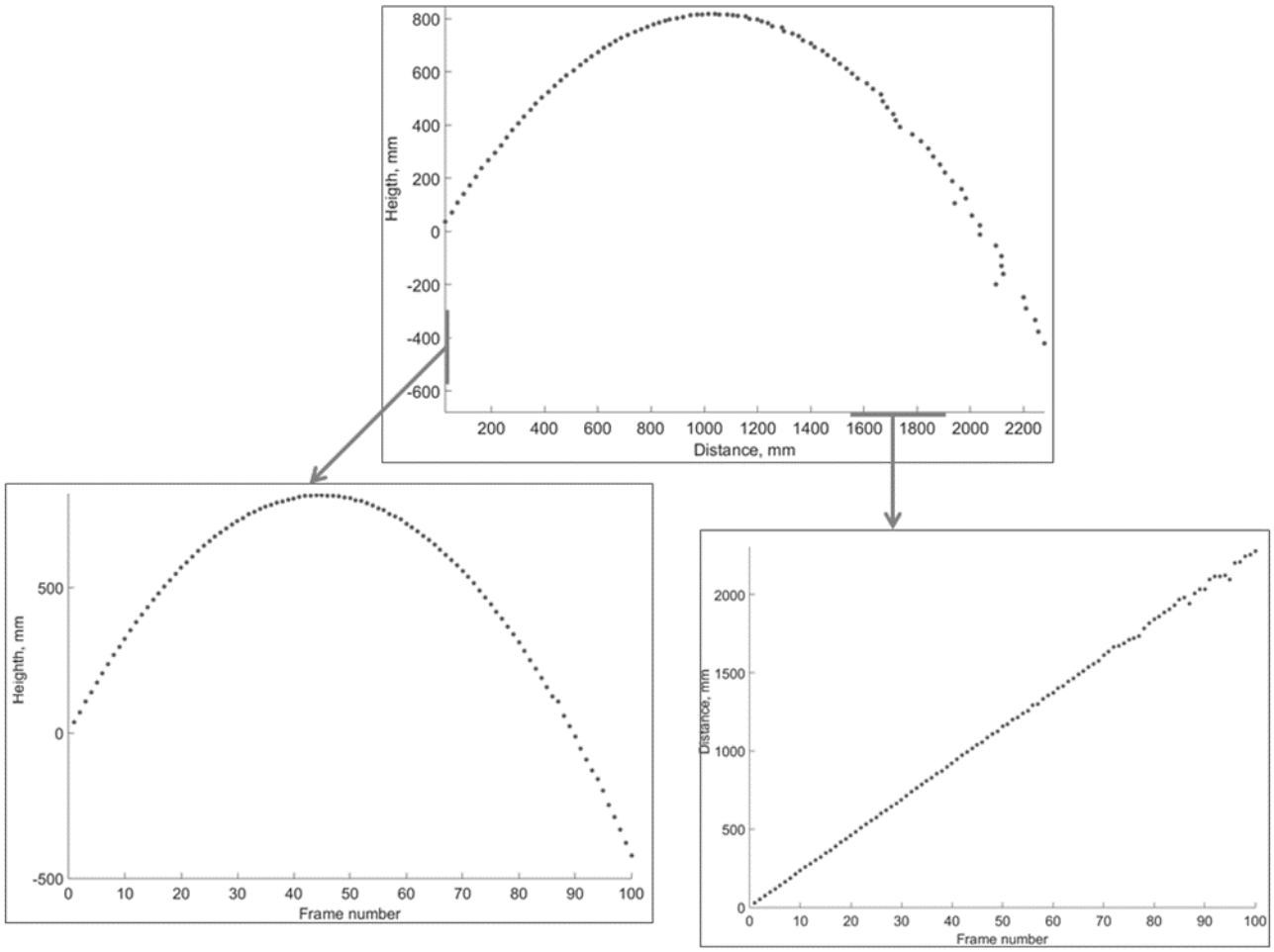


Figure 3. Dependence between height of the object and distance from the camera (upper graph), dependence of height on the frame number (bottom left) and dependence of the distance on the frame number (bottom right).

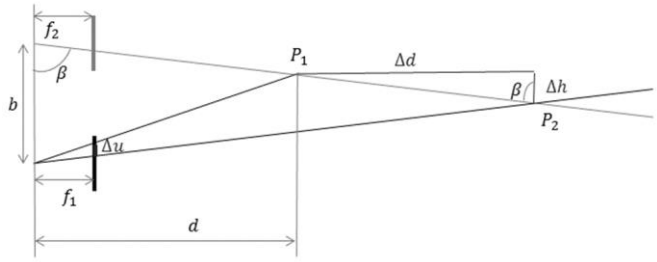


Figure 4. The effect of pixel error Δu on 3D positioning errors when measuring the height Δh and the distance Δd of an object.

stage is preferable than vice versa, due to the following factor. The final part of the trajectory is not processed in real time. Under actual transportation conditions, the ball will already be in the gripping workspace. Tracking the final part of the trajectory is used only for development of a trajectory training base; therefore, its accuracy can be improved by applying 1000 runs of RANSAC to the data. Accurate positioning is necessary to measure launch parameters: speed, throw angle, position in the first frame, etc. Another factor is ball capture; the measurement of the ball position in the final region will not be accurate in any case. The robot moving in the field of view generates excessive distortions in the functioning of the algorithm. Because of these factors, the location of the camera on the side of the thrower is more likely than vice versa. It would also be possible to arrange the cameras in a different way: at a greater distance from each other or not parallel to the direction of the trajectory. However, as a result, the measurement error will not be localized in dimension coinciding with the direction of the object motion, as shown in Figure 3. In fact, this localization is very useful for correcting errors. In this measurement, the object moves at an almost constant speed, and the movement can be approximated by a second-order polynomial.

Since it is undesirable to use analytical models of object movement, approximation is applied only for the distance from the camera to the object and only at the final stage of the trajectory (starting from the 60th frame). The graph of the measured and approximated values of the distance to the object is shown in Figure 5. From a visual point of view, the results of the approximation look believable.

The stereo system used to track the trajectory of a thrown object measures its position in the coordinate system associated with the optical center of the left camera. In principle, trajectory prediction can be made in this form as well, but then the degree of trajectory proximities will be determined not only by the similarity of their shape, but also by the direction of the throw and the position of the point from which the throw is made. We have proposed to transform the coordinates of the thrown object into such a system where the trajectories can be compared and predicted based solely on the shape of the trajectory.

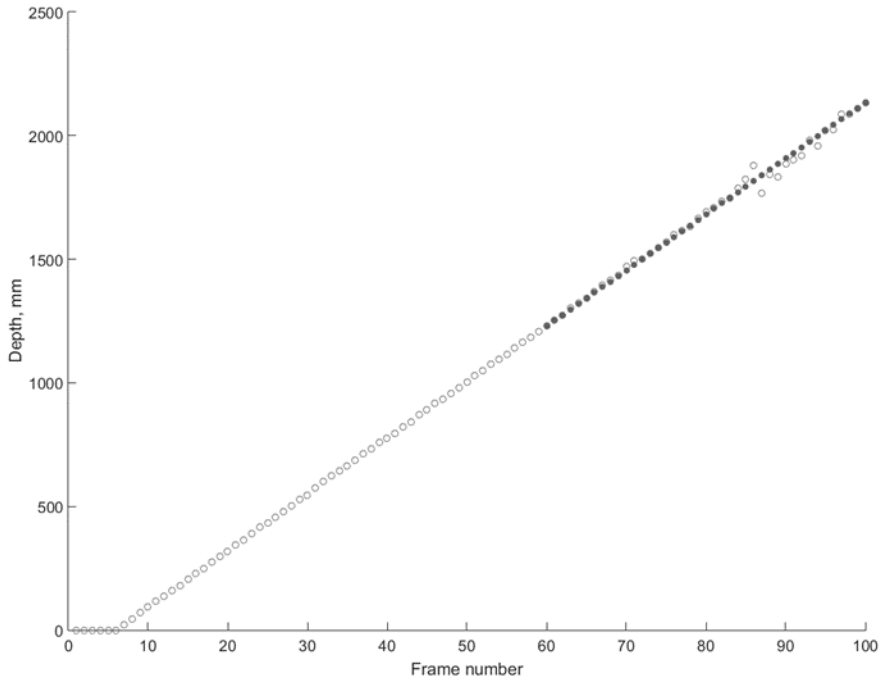


Figure 5. The difference between the measured (red circles) and approximated (blue points) values of the distance to the object

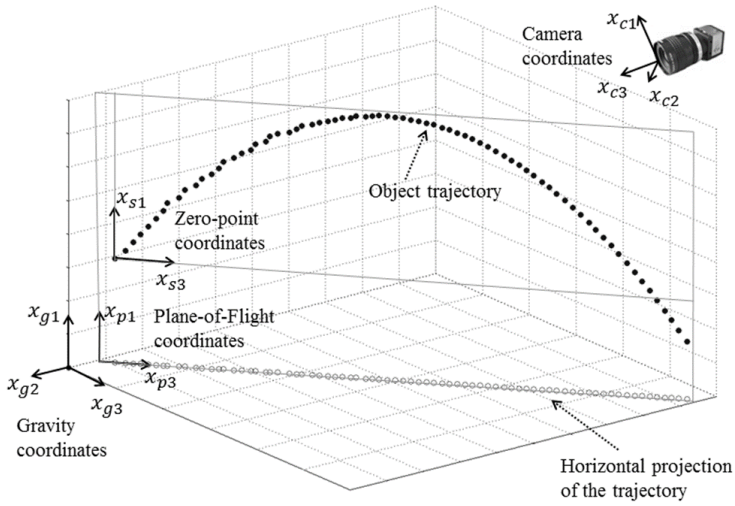


Figure 6. Mutual arrangement of coordinate systems

The purpose of coordinate transformations is to present the trajectories of thrown objects in a form in which it will be convenient to compare them. For this, the following sequence of coordinate transformations is proposed:- The three-dimensional system x_c1 Ox_c2 x_c3, in which the point O coincides with the optical center of the left stereo pair camera, and the axis x_c3 is aligned with the optical axis of the camera. In this coordinate system, the position of an object is measured by the stereo pair.

– The three-dimensional system x_g1 Ox_g2 x_g3, in which the axis x_g1 is aligned with the gravity vector, and the plane formed by the other two axes is, respectively, horizontal. This coordinate system allows you to localize the effect of gravity in one spatial dimension. As will be shown below, the transfer from such a system to the x_p1 Ox_p3 flight plane is simpler than from the x_c1 Ox_c2 x_c3 system. The transition matrix from x_c1 Ox_c2 x_c3 to x_g1 Ox_g2 x_g3 is determined during stereo system calibration. The gravity vector in the x_c1 Ox_c2 x_c3 system can be determined by hanging the load on the thread: in equilibrium, the thread is parallel to the desired vector.

– The two-dimensional x_p1 Ox_p3 system (flight plane), in which the x_p1 axis is aligned with the x_g1 axis, and the horizontal projection of the object velocity lies on the x_p3 axis. In the event that lateral forces do not act in flight on the body (they can be associated, for example, with the action of the wind or with the Magnus effect), the flight path lies in such a plane. Experiments conducted in [20] showed that the influence of lateral forces on the flight of an object can be neglected. Since the real direction of the throw in each case will be different from the others, then the transition matrix between the systems will be different for each case. The two-dimensional system x_s1 Ox_s3 in which the

directions of the axes coincide with those in x_p1 Ox_p3, and the center is located on one of the first points of the trajectory. Transferring the trajectory to such a coordinate system ensures result independence from the spatial location of the point from which the throw was made. An analysis of the measurement accuracy carried out in [28] showed that the accuracy in the first few frames is slightly worse than in the subsequent frames. Therefore, the sixth point on the trajectory was chosen as the common center of coordinates.

The mutual arrangement of the coordinate systems is shown in Figure 6. At the stage of predictor learning, all trajectories in the database are converted into the x_s1 Ox_s3 system and saved in this form. In the process of predictor's work, the current XC trajectory is converted to the x_s1 Ox_s3 system; prediction is performed in this system, and then the result YC is converted back to the original coordinate system.

This was the procedure of extracting information about tracking the thrown tennis ball. Considering more complex-shaped objects will require use of more specific image processing algorithm. The procedure of stereo triangulation will be the same, while the question of how to define pixel coordinates of the object's center must be answered by other means instead of circle recognition. Various methods for object positioning task were developed such as rule-based algorithms, pixel-based classification, analysis of brightness distribution, convolutional neural networks, and other techniques. Development of image processing approach for complex-shaped objects is a subject of future work.

3. Predicting the trajectory of a thrown object

From the point of view of the subtask of trajectory prediction, the existing systems of robotic object capture on the fly can be divided into three groups:

- Accurate throw systems. The high accuracy of the throw (that is, the small deviation of the initial velocity and direction of flight from the given value) makes it possible to ensure that the trajectories of the thrown objects turn out to be almost identical. In this case, there is no need to predict the trajectory anew after each throw. It is enough to make a throw once, to track the trajectory of the object and to develop the trajectory of the capture device based on the results. This approach is applicable to objects with high aerodynamic stability (for example, cylindrical objects in [16]). If dropped objects do not have the required aerodynamic stability (studies in [22] and [16] show that even objects that are as simple in shape as a tennis ball and a hollow metal cylinder, respectively, do not possess it), this approach ceases to be useful.

- Interactive capture systems. In such systems, prediction is not used: movement of the capture device is determined by the current position of the object. For example, in [6], the movement of the working body is set in such a way as to maintain a constant value of the angle of view for an object in an image from a camera attached to a capture device. In [5], at each moment in time, the movement of the robot is set in the direction of the current position of the object. The implementation of such systems requires a high response speed of a robotic capture device and a high efficiency of obtaining information about object movement (for example, in [5] a vision system was used, in which the frame rate was reduced to 1 kHz by directly connecting video matrix elements to the processor). The approach is not applicable if the throw is made from a long distance and you need to choose in which area of the working space to place the gripping device (just as the football player-goalkeeper first chooses which corner to defend and then catches the ball). Since transportation of objects in an industrial environment involves throwing over a distance of several meters, this method is not suitable for such systems.

- Systems with long-term forecasting. These include the ones described in [1-4,7-11] as well as the system discussed in this section. A more detailed overview of such systems is given below.

The majority of the ballistic trajectory predictors isbased on the modeling of forces acting on the body. In the simplest case, it is assumed that the only such force is gravity (as if the body was moving in a vacuum). Such a model was considered in [2,3,7]. Prediction of the trajectory is carried out by approximation of the measured values in a parabola. In [10], the model was extended to predict the trajectory of asymmetric objects (in the experiments, empty and half-filled plastic bottles, hammers, tennis rackets and boxes were used). Prediction was made on the basis of the assumption that the acceleration vector of the object was constant over all six degrees of freedom. Strictly speaking, this assumption is wrong. The body movement under the action of gravity and air resistance is given by a differential equation which has no analytical solution and is solved in practice by numerical methods. This approach was used to predict the trajectory in a number of works [4,8,9]. Further complication of analytical models leads to a significant increase in the volume of computation [15].

On the other hand, people acquire the ability to catch thrown objects at an early age without any knowledge of aerodynamics. We catch a thrown ball based on our previous experience. Because of this, it was suggested [18,19] to use a trajectory predictor based on previous experience. In [18], a neural network predictor was proposed as a means of prediction, but it did not provide adequate forecast accuracy. Moreover, the results of prediction by neural networks are difficult to interpret; therefore, it was later proposed to apply a more transparent method of k nearest neighbors [19]. The development of individual details of this method is described in the articles of [20,21].

Here we propose the method, which lies between model-based and learning-based model. The predictor is using equations to define future positions of the object, but the parameters of these equations are obtained by the learning procedure of genetic programming. Learning does not require a sampling of past trajectories: the parameters of the model are learned from the initial part of the current trajectory. Genetic programming (proposed by Cramer [36] and developed by Koza [37]) is not a synonym of genetic algorithm. Genetic programming is an application of the principles of genetic algorithms to automatic generation of a program code. In many applications including this research genetic programming is used for generating equations, which represent the process with unknown parameters. Target program (equation) is defined as a tree consisting of nodes and arcs. The nodes are operations and the arcs are operands. Initial versions of the tree are modified via the genetic operations (mutation, crossover,

selection), which are similar to the respective processes in genetics [37]. For the forecasting of trajectory, the task is to define the function for calculating future values of the coordinates based on its previous known coordinates. Genetic operations aim to define recurrent equation for trajectory forecasting. Genetic Programing OLS MATLAB toolbox was used to execute the genetic operations.

Table 2. Results of numerical experiments for 20 trajectories

#	*Equation*	*MSE, mm*
1	y(k-1) + (-0.038706) * (x(k-1)) + (0.033084)	6
2	y(k-1) + (-0.039329) * (x(k-1)) + (0.033553)	6
3	y(k-1) + (-0.038554) * (x(k-1)) + (0.031962)	5
4	y(k-1) + (-0.038689) * (x(k-1)) + (0.035645)	2
5	y(k-1) + (-0.038774) * (x(k-1)) + (0.033805)	7
6	y(k-1) + (-0.038546) * (x(k-1)) + (0.030245)	7
7	y(k-1) + (-0.038526) * (x(k-1)) + (0.032994)	4
8	y(k-1) + (-0.038740) * (x(k-1)) + (0.035160)	6
9	y(k-1) + (-0.038601) * (x(k-1)) + (0.033345)	3
10	y(k-1) + (-0.038454) * (x(k-1)) + (0.032473)	4
11	y(k-1) + (-0.038388) * (x(k-1)) + (0.031648)	6
12	y(k-1) + (-0.038501) * (x(k-1)) + (0.034544)	6
13	y(k-1) + (-0.038134) * (x(k-1)) + (0.033116)	9
14	y(k-1) + (-0.037313) * (x(k-1)) + (0.036494)	8
15	y(k-1) + (-0.038036) * (x(k-1)) + (0.032546)	4
16	y(k-1) + (-0.038347) * (x(k-1)) + (0.034728)	4
17	y(k-1) + (-0.038149) * (x(k-1)) + (0.034948)	3
18	y(k-1) + (-0.037972) * (x(k-1)) + (0.033884)	4
19	y(k-1) + (-0.037583) * (x(k-1)) + (0.032284)	7
20	y(k-1) + (-0.037841) * (x(k-1)) + (0.036221)	7

Numerical experiment on trajectory prediction was conducted using the tool [38] in two stages. On the first stage, we tried to define the common trajectory equations by learning from various trajectories. This try failed: the trajectories are different from each other and equation may be good for one trajectory and useless for another. Therefore we changed our strategy. On the second stage of experiment the initial part of each trajectory (first 50 frames) was used for learning the recurrent formula of this trajectory. Then the accuracy of this formula was checked on frames from 60 to 80. Learning procedure was applied to 20 trajectories acquired during the throwing experiments. The results are presented in table 2.

Each row shows the results for one trajectory. First column is trajectory number. The second one show equations, defining the height of the object y on frame number k as a function from its height y and distance x on previous frames. The right column show the standard deviation of the predicted ball position from the real one. It may be seen that the standard deviations of predicted height do not exceed 10 mm. According to the three-sigma rule the errors of prediction do not exceed 30 mm with high probability. All equations generated by the algorithm are relatively simple and have the same type: linear dependence from both coordinates of the previous frame. At the beginning of learning genetic operations often generate more complicated equations. These equations may include coordinate values from several previous frames.

4. Conclusion

We have proposed an algorithm for extraction and prediction of the ballistic trajectory based on video signal. Object trajectory is extracted from video sequence by the image processing algorithm, which include Canny edge detection, RANSAC circle recognition and stereo triangulation. The model of object motion is defined by the genetic programming. The result of the exploration is a recurrent formula for calculation of the object's position. Numerical experiments with real trajectories of the thrown tennis ball showed that the algorithm is able to forecast the trajectory accurately.

References

[1] Gayanov, R., Mironov, K., Kurennov D.: Estimating the trajectory of a thrown object from video signal with use of genetic programming, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, pp. 134 to 138, December 2017.

[2] Hove, B., Slotine, J.-J.: Experiments in Robotic Catching, American Control Conference, Boston, USA, pp. 381 to 386, 1991.

[3] Nishiwaki, K., Konno, A., Nagashima, K., Inaba, M., Inoue, H.: The Humanoid Saika that Catches a Thrown Ball, IEEE International Workshop on Robot and Human Communication, Sendai, Japan, pp. 94 to 99, October 1997.

[4] Frese, U., Baeuml, B., Haidacher, S., Schreiber, G., Schaefer, I., Haehnle, M., Hirzinger, G.: On-the-Shelf Vision for a Robotic Ball Catcher, IEEE/RSJ International Conference on Intelligent Robots and Systems, ,Maui, Hawaii, USA, pp. 591 to 596, November 2001.

[5] Namiki, A., Ishikawa, M.: Robotic Catching Using a Direct Mapping from Visual Information to Motor Command, IEEE International Conference on Robotics &Automation, Taipei, Taiwan, pp. 2400 to 2405, September 2003.

[6] Mori, R., Hashimoto, K., Miyazaki, F.: Tracking and Catching of 3D Flying Target based on GAG Strategy, IEEE International Conference on Robotics 8 Automation, New Orleans, USA, pp. 5189 to 5194, April 2004..

[7] Herrejon, R., Kagami, S., Hashimoto, K.: Position Based Visual Servoing for Catching a 3-D Flying Ob-ject Using RLS Trajectory Estimation from a Mo-nocular Image Sequence, IEEE International Conference on Robotics and Biomimetics, Guilin, China, pp. 665 to 670, December 2009.

[8] Baeuml, B., Wimboeck, T., Hirzinger, G.: Kinematically Optimal Catching a Flying Ball with a Hand-Arm-System, IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, pp. 2592 to 2599, October 2010.

[9] Baeuml, B., Birbach, O., Wimboeck, T., Frese, U., Dietrich, A., Hirzinger, G.: Catching Flying Balls with a Mobile Humanoid: System Overview and Design Considerations, IEEE-RAS International Conference on Humanoid Robots, Bled, Slovenia, pp. 513 to 520, October 2011.

[10] Kim, S., Shukla, A., Billard, A.: Catching Objects in Flight, IEEE Transactions on Robotics, Vol. 30, No. 5, pp. 1049 to 1065, May 2014.

[11] Cigliano P., Lippiello V., Ruggiero F., Siciliano B.: Robotic Ball Catching with an Eye-in-Hand Single-Camera System // IEEE Transactions on Control System Technology. 2015. Vol. 23. No. 5. P. 1657-1671.

[12] Frank, H., Wellerdick-Wojtasik, N., Hagebeuker, B., Novak, G., Mahlknecht, S.: Throwing Objects: a bio-inspired Approach for the Transportation of Parts, IEEE International Conference on Robotics and Biomimetics, Kunming, China, pp. 91 to 96, December 2006.

[13] Pongratz, M., Kupzog, F., Frank, H., Barteit, D.: Transport by Throwing - a bio-inspired Approach, IEEE International Conference on Industrial Informatics, Osaka, Japan, pp. 685 to 689, July 2010.

[14] Barteit, D.: Tracking of Thrown Objects, Dissertation, Faculty of Electrical Engineering, Vienna University of Technology, December 2011.

[15] Pongratz, M., Pollhammer, K., Szep, A.: KOROS Initiative: Automatized Throwing and Cathcing for Material Transportation, ISoLA 2011 Workshops, pp. 136 to 143, 2012.
[16] Frank, T., Janoske, U., Mittnacht, A., Schroedter, C.: Automated Throwing and Capturing of Cylinder-Shaped Objects, IEEE International Conference on Robotic and Automation, Saint Paul, Minnesota, USA, pp. 5264-5270, May 2012.
[17] Pongratz, M., Mironov, K. V., Bauer F.: A soft-catching strategy for transport by throwing and catching, Vestnik UGATU, Vol. 17, No. 6(59), pp. 28 to 32, December 2013.
[18] Mironov, K. V., Pongratz, M.: Applying neural networks for prediction of flying objects trajectory Vestnik UGATU, Vol. 17, No. 6(59) pp. 33 to 37, December 2013.
[19] Mironov, K.., Pongratz, M., Dietrich, D.: Predicting the Trajectory of a Flying Body Based on Weighted Nearest Neighbors, International Work-Conference on Time Series, Granada, Spain, pp. 699 to 710, June 2014.
[20] Mironov, K., Vladimirova, I., Pongratz, M.: Processing and Forecasting the Trajectory of a Thrown Object Measured by the Stereo Vision System, IFAC International Conference on Modelling, Identification and Control in Non-Linear Systems, St.-Petersburg, Russian Federation, June 2015.
[21] Mironov K. V., Pongratz M.: Fast kNN-based Prediction for the Trajectory of a Thrown Body, 24th Mediterranean Conference on Control and Automation MED 2016. Athens, Greece. pp. 512 to 517. July 2016.
[22] Mehta, R., Alam, F., Subic, A.: Review of tennis ball aerodynamics, Sports technology review, John Wiley and Sons Asia Pte Ltd, 2008, No. 1, pp. 7 to 16, January 2008.
[23] Szeliski, R.: Computer Vision: Algorithms and Applications, Springer, September 2010
[24] Zhang, Z.: A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, pp. 1330 to 1334, November 2000
[25] Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/, visited on September 18, 2018.
[26] Pongratz, M., Mironov, K. V.: Accuracy of Positioning Spherical Objects with Stereo Camera System, IEEE International Conference on Industrial Technology, Seville, Spain, pp. 1608 to 1612, March 2015.
[27] Mironov K. V.: Transport by robotic throwing and catching: Accurate stereo tracking of the spherical object, 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM).
[28] Lee, J.H., Akiyama, T., Hashimoto, H.: Study on Optimal Camera Arrangement for Positioning People in Intelligent Space, IEEE/RSJ international conference on intelligent robots and systems, Lausanne, Switzerland, pp. 220 to 225, October 2002.
[29] UI-3370CP – USB 3 Cameras – CAMERAFINDER – Products, http://en.ids-imaging.com/store/ui-3370cp.html, visited on September 25, 2014.
[30] Goetzinger, M.: Object Detection and Flightpath Prediction, Diploma Thesis, Faculty of Electrical Engineering, Vienna University of Technology, June 2015.
[31] Canny, J.: A Computational Approach to Edge De-tection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, November 1986.
[32] Hough, P.: A method and means for recognizing complex patterns, U.S. Patent No. 3,069,654, December 1962.
[33] Scaramuzza, D., Pagnotelli, S., Valligi, P.: Ball Detection and Predictive Ball Following Based on a Stereo-scopic Vision System, IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp. 1573-1578, April 2005.
[34] Fischler, M. A., Bolles, R. C.: Random Sample Con-sensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the Association for Computing Machinery, Vol. 24, No. 6, pp. 381 to 395, June 1981.
[35] Hlawatsch, F.: Parameter Estimation Methods: Lecture Notes, Grafisches Zentrum HTU GmbH, Vienna, Austria, March 2012.
[36] Cramer, N. A.: A Representation for the Adaptive Generation of Simple Sequential Programs, International Conference on Genetic Algorithms and their Applications, Pittsburgh, USA, pp. 183 to 187, July 1985.
[37] Koza, J.R.: Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Stanford University Computer Science Department technical report STAN-CS-90-1314, June 1990.
[38] GP-OLS MATLAB Toolbox [WebSite]. – http://www.abonyilab.com/software-and-data/gp_index/gpols, vizited on 04.09.2017.

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 251-259 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A New Model of Supplier's Selection for Sustainable Supply Chain Management

Oussama El Mariouli*,Abdellah Abouabdellah

Systems Engineering Laboratory, MOSIL, ENSA, University Ibn Tofail, Kenitra, Morocco

ARTICLE INFO

Article history:
Received: 17 December, 2018
Accepted: 04 March, 2019
Online: 03 April, 2019

Keywords:
Supply Chain
Criteria
Sustainable development
Supplier selection

ABSTRACT

In the last years, industrial company began to understand the central role that plays sustainable supplier selection (SSS) process for obtaining a sustainable supply chain (SSC). This paper proposes a new multi-criterion decision-making (MCDM) model for managers, which makes it possible to select the best suppliers who respect the sustainable development (SD) concept for a SSC. We start our paper by selecting the most recent economic, environmental and social criterion used in researches. After, we develop a new model mathematic that take into consideration the SD criteria in sustainable supplier rank and selection. Finally, we present a case study for testing our model.

1. Introduction

Recently, most industrial companies adopt the strategy of integrating the SD policy into their supply chain (SC), from downstream (purchase) upstream (distribution) to remain competitive and open up to new international markets, this approach starts with the selection of suppliers who respect SD's dimensions: SSS is central to the management of an SSC [1,2].

The new strategic direction of companies towards the SD concept, has required new economic, environmental and social criteria instead of conventional criteria (quality, cost and lead time) in purchasing decisions, therefore the process of supplier selection is has become a complex and multi-criteria decision that depends on several qualitative and quantitative criteria.

In this paper, we develop a new multi-objective mathematical model that allows measuring the index of supplier's sustainability for chooses the best between them. Our research methodology is articulated around two phases of realization, which we will detail them below:

- The first phase is the identification of recent SD criteria used to measure supplier performance;
- Last Phase present the equation that measures and rank sustainable supplier performance.

2. Literature Review

This section has four subsections. The first part presents a brief overview of the SSC management. The second part presents work on selection criteria. The last part presents the different vendor selection models found in the literature search and the last section presents the methods and approaches used in ranking the best SS.

2.1. Sustainable supply chain management

In the last 10 years, the SSC has become an important topic in the businesses and academies [3]. The role of the SSCM is to manage flows, information and cooperation among organizations throughout the supply network, taking into account economic, environmental and social criteria [4].

2.2. Supplier selection criteria

In the literature, we noticed that the selection criteria of suppliers change with the globalization and the appearance of the new concept such as Just-a-temp.

The first selection criteria for suppliers were made by Dickson (1966); he conducted a survey of 274 Canadian and American companies that are members of the National Association of Purchasing Managers (NAMP), and identified 23 criteria used by companies in the 1960s to select their suppliers [5].

The study by Weber and al. (1991) showed that these criteria remained the same until the 1990s, but the importance and the coefficient relative to each of the criteria are changed [6].

The study by Larson (1993) showed that the quality and the total cost of the product purchased are the most important criteria in the selection of suppliers [7].

The study by Vonderembse and al. (1995) out of 268 US companies, NAPM member has defined 10 important criteria used

*Oussama El Mariouli, Email: oussama.el.mariouli@gmail.com

for the choice of suppliers; it has shown that these companies specify the performance and the quality of the product in the selection of the suppliers [8].

The study by Verma and al. (1998) on 323 US companies showed 4 criteria (quality, price, delay and flexibility) for the selection of suppliers [9].

The study by Katsikeas and al. (2004) out of 237 British companies showed that the choice of suppliers is based on 4 criteria (reliability of the delays, price competitiveness, service offered and technological capacity) [10].

The study by Aguezzoul and Ladet (2006) showed that the QCD triptych (quality, cost and delay) was the most used in the selection of suppliers [11].

The Baumann study (2011) defined 15 indicators to measure the sustainable development of the supply chain [12].

The study by Boukherroub and al (2013), inspired by Baumann (2011), defines 7 criteria for sustainable development because of their relevance in the scientific literature and international standards [13].

The study of Ouzlem and al. (2015) proposed a list formed by 8 main criteria and 31 ecological subscribers [14].

The study of Caron and al. (2016) identify 8 principles and 27 criteria of SD using in mineral exploration industry [15].

The study of Song and al. (2017) used 10 criteria to SSS for solar air-conditioner manufacturer [16].

The study of Vasiljević and all. (2018) proposed 20 criteria used to evaluate the supplier in automotive industry [17].

The study of Sureeyatanapas and al. (2018) proposed 39 supplier selection criteria [18]. The study of Memari and al. (2019) presented the top 10 economic, environmental and social criteria for SSS [19].

In the second section, we will give a new list of recent criteria used for the supplier selection in a context of sustainable development.

2.3. Selection of suppliers

Supplier selection is one of the most crucial decisions made by managers in organizations [20]. The decision has become multi-criteria and very complex, it depends on qualitative and quantitative factors at the same time. Several research and literature reviews [21-23, 11] presented different methods and approaches for help buyers make the decision to select the best supplier. The table (Table1) presents the different methods of selection and evaluation of suppliers found in the literature.

In the literature, we find several recent researches that study the problem of the selection of suppliers. We can notice that:

- Most of this research does not include the three sustainable development dimensions in the supplier selection process; it focused only on green supplier problem [24-30].
- Most methods developed are complex to use.
- AHP is the most used method in the literature, but this method has the disadvantage of uncertainty of judgment.

In this study, our model developed has as originality:

- Integration of SD criteria most used in literature and internationals standards in the supplier selection process.
- Simple to apply and can be used in any industrial enterprise.

Table 1: Methods of selection and evaluation of suppliers

Supplier selection method		***Disadvantage***
Artificial intelligence	**CBR** (Case Based Reasoning system)	- The collection and processing of supplier data by experts take a lot of time.
	ES (Expert System).	
	NN (Neural Network)	
	FST (Fuzzy Sets Theory)	
Statistical / Probabilistic Models	**CA** (Cluster Analysis)	- Absence of optimal solution - Difficulty of analysis - Absence of the possibility of adding mathematical constraints in the model
	FA (Factor Analysis)	
Model of mathematical programming	**DEA** (Data Envelopment Analysis)	- Difficulty of dealing with subjective criteria - Absence of optimal solution - Difficulty analyzing the result obtained from the method
	GP (Goal Programming) or multi-objective programming	
	MOP (Multi Objectifs Pprogramming)	
	AHP (Analytic Hierarchy Process)	
	FST (Fuzzy Sets Theory)	
	ANP (Analytic Network Process)	
	TOPSIS (Technique for the Order Performance by Similarity to Ideal Solution)	
	MAUT (Multiple Attribute Utility Theory)	
	TCO (Total Cost Ownership)	
The categorical method	It relies on the experience and expertise of the buyer	- A subjective method - Absence of the relative importance of each criterion
Hybrid models	It represents the combination of two or more methods of evaluation	- Computational complexity

3. Methodology

3.1. Determining criteria for measuring sustainable supplier performance

This paper aims to provide a broad synthetic analysis of an area that has experienced rapid growth in knowledge [40]. The data collected for this paper will help to identify the SD criteria that are cited in the various published research to define a useful theoretical overview to better understand SS phenomena.

The literature review follows a systematic methodology (Briner and Denyer 2012) to reduce the number of articles published and focuses on leading journals that publish SS research [41]. In our initial library research on relevant documents. we did an analysis of the different databases "ScienceDirect, Springer, Emerald, Taylor Francis, Inderscience, GoogleShooler ..." by the combination of keywords "procurement, supply chain, sustainable development, performance, indicators, supplier's selection, review of the literature, criteria, economiec, social, environment", These words were carefully chosen to ensure that as many relevant articles as possible could be included.

In the literature, we identified the first study of supplier selection criteria was in 1966 (Dickson 1966), we followed the documents until 2018. Also we searched at international standards such as SCOR, GRI, OECD sustainable toolkit [31] and ISO 26000...

We have observed an increase in the work published from the year 1966 to the beginning of 2018.

The filtration process was carried out in three steps. In the first step, we found an initial list of 168 criteria that are related to SD, after we chose 34 criteria that are most relevant in the literature and finally, we defined 28 criteria that are related to the SSS process (fig.1).

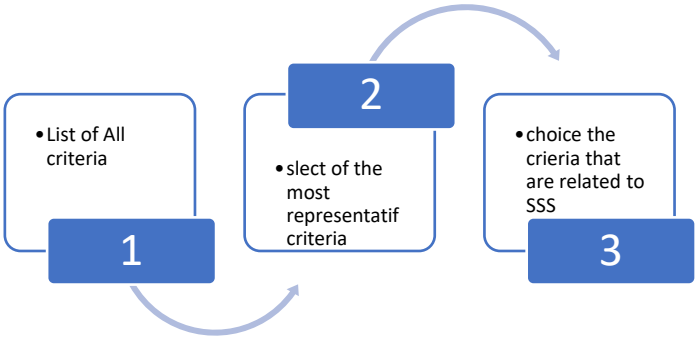


Figure 1: Follow-up methodology

A review of the literature and a database search were carried out to build a common list of SD criteria for the SS process and in international standards, as a result:

Economic criteria

The traditional model of supplier selection depended on economic criteria (quality, cost and time), but with increasing concern about environmental and social issues, other criteria are added in the purchasing decision. The economic criteria found in the literature for the selection of suppliers are:

- Innovation capacity (C_1): Is the ability to bring new solutions or to improve an existing solution in order to be competitive.
- Production capacity (C_2): The ability to meet customer needs.
- Technical and technological capacity (C_3): The use of new techniques and technologies in production.
- Cost (C_4): This is the cost associated with purchasing the raw material or component provided.
- Deadlines (C_5): This is the time needed to respond to an expressed need.
- Reliability (C_6): Customer satisfaction in terms of product choice, quality or lead time.
- Financial (C_7): Costs related to the design and/or development of products.
- Flexibility (C_8): Is the ability to respond to situations and likely disruptions.
- Delivery (C_9): Transport of goods and assurance of their delivery to the destination.
- Quality (C_{10}): It is defined by standards such as ISO to satisfy expressed needs.
- Reactivity (C_{11}): It is the ability to respond quickly to increasingly diverse needs in the global market.
- Customer references (C_{12}): It is the development of references, testimonials and customer success.

Environmental criteria

Environmental criteria are aimed at making choices of products and services that minimize the exploitation and consumption of natural resources, avoid the production of waste and reduce the risk of contamination of the living environments arising from consumer habits and the whole Activities related to their life cycle. The environmental criteria found in the literature for the selection of suppliers are:

- Waste (C_{13}): This is the waste from manufacturing.
- Emission (C_{14}): These are the various gases and substances emitted from the manufacture and transport of products.
- Environmental label (C_{15}): Quality labels to establish that if a product or actor has a reduced impact on the environment.
- Pollution (C_{16}): Pollution emitted in air, soil and water due to production.
- Program (C_{17}): These are the different programs and management policy for the protection of the environment.
- Recycling (C_{18}): The reuse of raw materials.

- Compliance with rules (C_{19}): Respect for environmental and social requirements.
- Use of resources (C_{20}): Materials used in the component supplied and the energy consumption during the manufacture of the product.
- Toxic or dangerous substances (C_{21}): The use of hazardous materials in the manufacture of products.

Social criteria

Social criteria support the ethical management of staff at CS suppliers, equity and encourage the responsibility of suppliers. Social criteria are little discussed in the literature; research in the literature has resulted in:

- Human rights (C_{22}): Are represented by the absence of forced labor and child labor, freedom of association and non-discrimination.
- Jobs and wealth (C_{23}): Improving the standard of living by ensuring full employment and stable employment.
- Training, support and education (C_{24}): All continuing training to ensure the development of human resources.
- Occupational health and safety (C_{25}): This is the protection of workers against the risks and adaptation of the working environment to the physiological and psychological needs of workers.
- Working conditions (C_{26}): Improvement of wages and respect for working time, rest periods, prayer periods, holidays, disciplinary practices and dismissals, maternity protection issues and such as access to drinking water, canteens and access to medical services.

3.2. Proposed mathematical model for measuring supplier sustainability index

We consider a manger that would like to select the best supplier who respects the three dimensions of SD. The following mathematical model allows decision-makers to choose the most sustainable supplier:

Parameter

N_{ns}	Number of new solution
N_{is}	Number of improved solution
Q_s	Quantity served
Q_d	Quantity requested
N_{nt}	Number of techniques and technologies in production renewed
C_{ost}	Cost associated with the purchase of the raw material or component provides (USD)
N_d	Number of days needed to respond to a expressed need
N_{cs}	Number of satisfied customers
N_{tc}	Total number of customers
C_{cd}	Costs related to product design and / or development (USD)
I_{fl}	Index ability to react to situations and likely disturbances
C_{ostd}	Cost of transporting goods to destination (USD)
N_s	Number of quality standards
I_{re}	Index of ability to respond very quickly to increasingly diverse needs in the global market
N_{gt}	Number of good testimonials
N_{tt}	Total number of testimonials
I_{wm}	Waste quantity index of product manufacturing
I_{egs}	Quantity index of the different gases and substances emitted from the manufacture and transport of products
I_{ql}	Quality Label Index
I_{pe}	Pollution index of air, soil and water due to production
I_{pp}	Index of different programs and management policy for the protection of the environment
I_{rm}	Index of reuse of raw materials
I_{rr}	Index of compliance with environmental and social requirements
I_{ur}	Index of quantity of materials (lead, chromium, coal ...) and energy (gas, water, electricity ...) used during the production of the product
I_{tds}	Quantity index of hazardous or toxic materials in the product
I_{hr}	Index of respect for human rights
I_{jw}	Improvement index for the standard of living of employers
I_{tse}	Index of continuing education to ensure the development of human resources
N_{wa}	Number of employees injured
N_{tw}	Total number of employees
I_{cw}	Index of respect for working conditions
I_{eco}	Index of economic dimension

I_{env} Index of environmental dimension

I_{soc} Index of social dimension

W_i Weight associated with the criterion C_i

W_{eco} Weight associated with the economic dimension

W_{env} Weight associated with the environmental dimension

W_{soc} Weight associated with the social dimension

Objective function

In this paper, the objective function is the index of sustainability supplier. It is represented by set of SD criteria having a value limited between 1 and 0. We use the ratio for facilitating the aggregation of the different elements in the objective.

The following equations represent the calculation of the set of SD criteria:

$$C_1 = \frac{N_{ns} + N_{is}}{N_{ns} + N_{is} + 1} \quad (1)$$

$$C_3 = \frac{Q_S}{Q_D} \quad (2)$$

$$C_3 = \frac{N_{nt}}{N_{nt} + 1} \quad (3)$$

$$C_4 = \frac{1}{C_{ost}} \quad (4)$$

$$C_5 = \frac{1}{N_d} \quad (5)$$

$$C_6 = \frac{N_{cs}}{N_{tc}} \quad (6)$$

$$C_7 = \frac{C_{cd}}{C_{cd} + 1} \quad (7)$$

$$C_8 = \frac{1}{I_{fl}} \quad (8)$$

$$C_9 = \frac{1}{C_{ostd}} \quad (9)$$

$$C_{10} = \frac{N_s}{N_s + 1} \quad (10)$$

$$C_{11} = \frac{1}{I_{re}} \quad , \quad I_{re} = \begin{cases} 1 \text{ if Good} \\ 2 \text{ if Average} \end{cases} \quad (11)$$

$$C_{12} = \frac{N_{gt}}{N_{tt}} \quad (12)$$

$$C_{13} = \frac{1}{I_{wm}} \quad , \quad I_{wm} = \begin{cases} 1 \text{ if Little} \\ 2 \text{ if Average} \end{cases} \quad (13)$$

$$C_{14} = \frac{1}{I_{egs}} \quad , \quad I_{egs} = \begin{cases} 1 \text{ if Little} \\ 2 \text{ if Average} \end{cases} \quad (14)$$

$$C_{15} = I_{ql} \, , \; I_{ql} = \begin{cases} 1 \text{ if the product has a quality label} \\ 0 \text{ if not} \end{cases} \quad (15)$$

$$C_{16} = \frac{1}{I_{pe}}, \quad I_{pe} = \begin{cases} 1 \text{ if Little} \\ 2 \text{ if Average} \\ 3 \text{ if Much} \end{cases} \quad (16)$$

$$C_{17} = \frac{1}{I_{pp}} \quad , \quad I_{pp} = \begin{cases} 1 \text{ if Beaucoup} \\ 2 \text{ if Average} \\ 3 \text{ if Little} \end{cases} \quad (17)$$

$$C_{18} = I_{rm} \quad , I_{rm} = \begin{cases} 1 \text{ if Recyclable} \\ 0 \text{ if No} \end{cases} \quad (18)$$

$$C_{19} = \frac{1}{I_{rr}} \quad , \quad I_{rr} = \begin{cases} 1 \text{ if Much} \\ 2 \text{ if Average} \\ 3 \text{ if Little} \end{cases} \quad (19)$$

$$C_{20} = \frac{1}{I_{ur}} \quad , \quad I_{ur} = \begin{cases} 1 \text{ if Little} \\ 2 \text{ if Average} \\ 3 \text{ if Much} \end{cases} \tag{20}$$

$$C_{21} = I_{tds} \quad , I_{rm} = \begin{cases} 1 \text{ if Recyclable} \\ 0 \text{ if No} \end{cases} \tag{21}$$

$$C_{22} = \frac{1}{I_{hr}} \quad , I_{hr} = \begin{cases} 1 \text{ if Good} \\ 2 \text{ if Average} \end{cases} \tag{22}$$

$$C_{23} = \frac{1}{I_{jw}} \quad , I_{jw} = \begin{cases} 1 \text{ if Good} \\ 2 \text{ if Average} \end{cases} \tag{23}$$

$$C_{24} = \frac{1}{I_{tse}} \quad , I_{tse} = \begin{cases} 1 \text{ if Much} \\ 2 \text{ if Average} \end{cases} \tag{24}$$

$$C_{25} = \frac{N_{wa}}{N_{tw}} \tag{25}$$

$$C_{26} = \frac{1}{I_{cw}} \quad , I_{cw} = \begin{cases} 1 \text{ if Good} \\ 2 \text{ if Average} \end{cases} \tag{26}$$

After obtaining the criteria values, decision makers are asked to assign weights to the chosen criteria. Each criteria value is multiplied by its weight. We obtain three value represents economic, environmental and social indicator, respectively as follows:

$$I_{eco} = \frac{\sum_{i=1}^{12} W_i C_i}{12} \qquad i = \{1,2,\ldots,12\} \tag{27}$$

$$I_{env} = \frac{\sum_{i=13}^{21} W_i C_i}{9} \qquad i = \{13,14,\ldots,21\} \tag{28}$$

$$I_{soc} = \frac{\sum_{i=22}^{26} W_i C_i}{5} \qquad i = \{22,23,\ldots,26\} \tag{29}$$

In the end, the decision-makers are asked to choose the weights of the three dimensions of sustainable development according to their policy followed by the company, then each value of SD dimension indicator obtained by equation (27) to (29) are multiplied by these weight.

The sustainability performance score of the supplier is given by the following maximum function:

$$\text{Max } P = \frac{W_{eco} I_{eco} + W_{env} I_{env} + W_{soc} I_{soc}}{3} \tag{30}$$

The MaxP equation represents the function of the developed model. The maximum value of the equation (30) is one.

Constraints:

- The sum of the weights in each dimension must not exceed the value 1.

$$Q_S \le Q_D \tag{31}$$

- The sum of the weights in each dimension must not exceed the value 1.

$$\sum_{i=1}^{12} W_i = 1 \tag{32}$$

$$\sum_{i=13}^{21} W_i = 1 \tag{33}$$

$$\sum_{i=22}^{26} W_i = 1 \tag{34}$$

$$W_{eco} + W_{env} + W_{soc} = 1 \tag{35}$$

- The sum of the weights of the dimensions of SD is equal to 1.

$$0 \le C_i \le 1 \qquad \forall i = \{1,2,\ldots,26\} \tag{36}$$

4. Data analysis

A company X decides to choose the best supplier among five suppliers. Managers have chosen the weights of each criterion in such a way that: the sum of the weights of the criteria of each dimension is equal to one.

All criteria value under each of the three pillars are multiplied by their respective weights and aggregated as indicated by Equations (27) to (29). The result is three index

Table 2: Index of environmental, economic and social dimensions

Dimension	Criteria	Weight	Supplier1	Supplier2	Supplier3	Supplier4	Supplier5
Economic	C_1	0,101	0,500	0,600	0,700	0,800	0,900
	C_2	0,079	0,200	0,400	0,600	0,800	1,000
	C_3	0,106	0,500	0,600	0,700	0,800	0,900
	C_4	0,115	0,910	0,920	0,930	0,940	0,950
	C_5	0,095	0,167	0,200	0,250	0,333	0,500
	C_6	0,098	0,500	0,600	0,700	0,800	0,900
	C_7	0,061	0,917	0,929	0,941	0,950	0,980
	C_8	0,053	0,333	0,500	0,500	1,000	1,000
	C_9	0,051	0,005	0,006	0,007	0,010	0,011
	C_{10}	0,082	0,710	0,770	0,790	0,870	0,920
	C_{11}	0,079	0,500	0,500	1,000	1,000	1,000
	C_{12}	0,080	0,850	0,880	0,900	0,980	1,000
	I_{eco}		0,528376	0,595915	0,691638	0,701738	0,858031
Environment	C_{13}	0,112	0,333	0,500	0,500	1,000	1,000
	C_{14}	0,115	0,333	0,333	0,500	0,500	1,000
	C_{15}	0,133	0,000	0,000	1,000	1,000	1,000
	C_{16}	0,137	0,333	0,333	0,500	0,500	1,000
	C_{17}	0,130	0,333	0,333	0,500	0,500	1,000
	C_{18}	0,074	0,000	0,000	1,000	1,000	1,000
	C_{19}	0,154	0,333	0,333	0,500	0,500	0,500
	C_{20}	0, 074	0,333	0,333	0,333	0,333	0,500
	C_{21}	0, 071	0	0	0	1,000	1,000
	I_{env}		0,240426	0,25913	0,555642	0,682642	0,886
Social	C_{22}	0,232	0,333	0,333	0,500	0,500	1,000
	C_{23}	0,129	0,333	0,333	0,500	0,500	1,000
	C_{24}	0,198	0,333	0,333	0,500	0,500	1,000
	C_{25}	0,191	0,200	0,240	0,250	0,290	0,300
	C_{26}	0,250	0,5	0,5	1,000	1,000	1,000
	I_{soc}		0,349347	0,356987	0,57725	0,58489	0,8663

values representing the environmental, economic and social dimensions, respectively (Table2).

The enterprise X choice the scenario of weigh dimension like as:

$$W_{eco} = W_{env} = W_{soc} = \frac{1}{3}$$

The final sustainable performance indices of suppliers were used to rank the suppliers in the X enterprise (Table3), supplier5 is identified as the best sustainable supplier with MaxP=0,29 (Fig2).

5. Conclusion

With the appearance of the new sustainable development concept in the last 20 years, industrial company have begun to

Table 3: The sustainable performance indices of the supplier

	I_{eco}	I_{env}	I_{soc}	sustainable performance	Rank
Supplier1	0,528376	0,240426	0,349347	0,12423878	5
Supplier2	0,595915	0,25913	0,356987	0,13467022	4
Supplier3	0,691638	0,555642	0,57725	0,20272556	3
Supplier4	0,701738	0,682642	0,58489	0,21880778	2
Supplier5	0,858031	0,886	0,8663	0,29003678	1

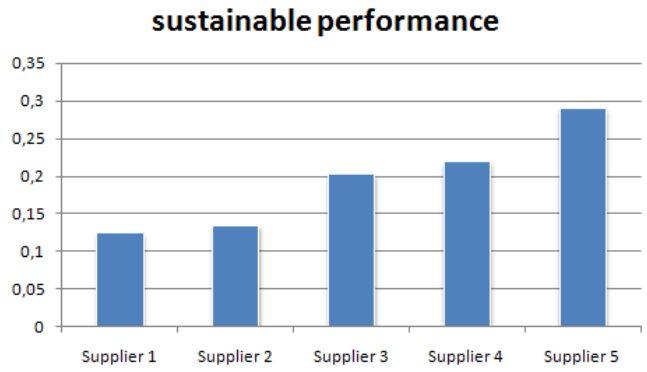


Figure 2: Sustainable performance of the supplier

integrate environmental and social concerns into the management of their activities in order to protect the company's durability and open up new markets. The process of integrating SD into the company starts with the selection of suppliers. In this paper, we have given a list that includes all the relevant criteria in the literature for the selection of suppliers. These criteria influence the functioning of all the other services of the company (transport, stock, production, etc.). In the last, we have presented an easy-to-use mathematical model for managers to measure vendor performance. Our developed model is simple to apply and can be used in any industrial enterprise to select the best sustainable supplier.

References

[1] Atefeh Amindousta, Shamsuddin Ahmeda, Ali Saghafiniab, Ardeshir Bahreininejada, Sustainable supplier selection: A ranking model based on fuzzy inference system, Applied Soft Computing 12 (2012) 1668–1677.

[2] Buyuközkan, G., Çifci, G., 2011. A novel fuzzy multi-criteria decision framework for sustainable supplier selection with incomplete information. Comput. Ind. 62, 164–174.

[3] X. Zhou, et al., Type-2 fuzzy multi-objective DEA model: An application to sustainable supplier evaluation, Appl. Soft Comput. J. (2016), http://dx.doi.org/10.1016/j.asoc.2016.04.038.

[4] Fallahpour, A., Udoncy Olugu, E., Nurmaya Musa, S., Yew Wong, K., Noori, S., A decision support model for sustainable supplier selection in sustainable supply chain management, Computers & Industrial Engineering (2017), doi: http://dx.doi.org/10.1016/j.cie.2017.01.005.

[5] Dickson, G. W. (1966). An analysis of vendor selection systems and decisions. Journal of Purchasing, 2 (1), 28-41.

[6] Weber, C. A., Current, J. R., & W. C. Benton (1991). Vendor selection criteria and methods. European Journal of Operational Research, 50, 2-18.

[7] Larson, P. D. (1993). Buyer-Supplier co-operation, product quality and total cost. International Journal of Physical Distribution & Logistics Management, 24 (6), 4-10.

[8] Vonderembse, M., Tracey, M., Tan, C. L., & E. J. Bardi (1995). Current purchasing practices and JIT: some of effects on inbound logistics. International Journal of Physical Distribution & Logistics Management, 25 (3), 33-48.

[9] Verma, R., & M. E. Pullma (1998). An analysis of the supplier selection process. International Journal of Management Science, 26 (6), 739-750.

[10] Katsikeas, C. S., Paparoidamis, N. G., & E. Katsikea (2004). Supply source selection criteria: the impact of supplier performance on distributor performance. Indutrial Marketing Management.

[11] Aicha Aguezzoul, Pierre Ladet. S_election et _evaluation des fournisseurs : Crit_eres et m_ethodes. Revue fran_caise de gestion industrielle, Paris-La-D_efense : Association fran_caise de gestion industrielle ; Montrouge : Centrale des revues, 2006, 2, pp.5-27.

[12] Emilie Baumann. Modèles d'évaluation des performances économique, environnementale et sociale dans les chaines logistiques. Gestion et management. INSA de Lyon, (2011).

[13] Tasseda Boukherroub. Int_egration des objectifs du d_eveloppement durable dans la gestion strat_egique et tactique de la chaine logistique. Gestion et management. INSA de Lyon, 2013.

[14] Ozlem Gurel, A. Z. Acar, I. Onden, I. Gumus (2015). Determinants of the green supplier selection. Procedia - Social and Behavioral Sciences 181 (2015) 131 – 139.

[15] Joanie Caron, Suzanne Durand, Hugo Asselin, (2016). Principles and criteria of sustainable development for the mineral exploration industry. Journal of Cleaner Production 119 (2016) 215-222.

[16] Wenyan Song, Zhitao X, Hu-Chen Liu, (2017). Developing sustainable supplier selection criteria for solar air-conditioner manufacturer: An integrated approach. Renewable and Sustainable Energy Reviews 79 (2017) 1461–1471.

[17] Marko Vasiljević, Hamed Fazlollahtabar, Željko Stevića, Slavko Vesković, (2018). A ROUGH MULTICRITERIA APPROACH FOR EVALUATION OF THE SUPPLIER CRITERIA IN AUTOMOTIVE INDUSTRY. Decision Making: Applications in Management and Engineering Vol. 1, Number 1, 2018, pp. 82-96. ISSN: 2560-6018. DOI: 10.31181/dmame180182v.

[18] Panitas Sureeyatanapas, Kawinpob Sriwattananusart, Thanawath Niyamosothath, Weerapat Setsomboon, Sirawadee Arunyanart, Supplier selection towards uncertain and unavailable information: an extension of TOPSIS method, Operations Research Perspectives (2018), doi: 10.1016/j.orp.2018.01.005.

[19] Ashkan Memaria, Ahmad Dargib, Mohammad Reza Akbari Jokara, Robiah Ahmadc, Abd. Rahman Abdul Rahimc, (2019). Sustainable supplier selection: A multi-criteria intuitionistic fuzzy TOPSIS Method. Journal of Manufacturing Systems 50 (2019) 9–24.

[20] Chunguang Bai, Joseph Sarkis, (2010). Integrating sustainability into supplier selection with grey system and rough set methodologies. Int. J. Production Economics 124 (2010) 252–264.

[21] Hacer Guner Goren, A Decision Framework for Sustainable Supplier Selection and Order Allocation with Lost Sales, Journal of Cleaner Production (2018), doi: 10.1016/j.jclepro.2018.02.211.

[22] Konrad Zimmer, Magnus Fröhling & Frank Schultmann (2015): Sustainable supplier management – a review of models supporting sustainable supplier selection, monitoring and development, International Journal of Production Research, DOI: 10.1080/00207543.2015.1079340.

[23] Ilgin MA, et al. Use of MCDM techniques in environmentally conscious manufacturing and product recovery: State of the art. J Manuf Syst (2015), http://dx.doi.org/10.1016/j.jmsy.2015.04.010.

[24] Babbar C, Amin SH. A multi-objective mathematical model integrating environmental concerns for supplier selection and order allocation based on fuzzy QFD in beverages industry. Expert Syst Appl 2018;92:27 – 38

[25] Jain, Vipul, Kumar Sameer, Kumar Amit, Chandra Charu. An integrated buyer initiated decision-making process for green supplier selection. J Manuf Syst 2016;41:256 – 65.

[26] Kannan D, Govindan K, Rajendran S. Fuzzy Axiomatic Design approach based green supplier selection: a case study from Singapore. J Clean Prod 2015;96:194 – 208.
[27] Ghayebloo S, Tarokh MJ, Venkatadri U, Diallo C. Developing a bi-objective model of the closed-loop supply chain network with green supplier selection and disassembly of products: the impact of parts reliability and product greenness on the recovery network. J Manuf Syst 2015;36:76 – 86.
[28] Govindan K, Rajendran S, Sarkis J, Murugesan P. Multi criteria decision making approaches for green supplier evaluation and selection: a literature review. J Clean Prod 2015;98:66 – 83.
[29] Banaeian, N., Mobli, H., Fahimnia, B., Nielsen, I.E., and Omid, M. (2018). Green supplier selection using fuzzy group decision making methods: A case study from the agri-food industry. Computers & Operations Research 89, 337-347.
[30] Keshavarz Ghorabaee, M., Zavadskas, E.K., Amiri, M., and Esmaeili, A. (2016). Multi-criteria evaluation of green suppliers using an extended WASPAS method with interval type-2 fuzzy sets. Journal of Cleaner Production 137, 213-229
[31] OECD Sustainable Manufacturing Toolkit. Avalable online: http://www.oecd.org/sti/inno/ 48101937.pdf (accessed on 25 December 2014).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 252-258 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

3D Reconstruction of Monuments from Drone Photographs Based on The Spatial Reconstruction of The Photogrammetric Method

Andras Molnar*

John von Neumann faculty of Informatic, Obuda University, 1034 Budapest, Hungary

ARTICLE INFO

Article history:
Received: 25 September, 2018
Accepted: 15 October, 2018
Online: 10 November, 2018

Keywords:
Drone
Aerial Photography
3d Surface Model
Photogrammetry,
Monument Protection
Virtual Museum

ABSTRACT

Due to their efficient flight control systems and their camera of high quality modern drones can fly precisely and take aerial photos of high resolution. Although these multi-rotor devices are not able to fly long distances yet, they are very efficient instruments for taking aerial photographs of their proximate environment. As they are able to float and rotate along their vertical axis to a discretionary direction, they can be used for monitoring and taking pictures of a building from several directions. If pictures are taken of a building from all directions with significant overlap (of at least 50%) the 3D reconstruction of the building in a photogrammetric way becomes possible. The reconstructed models do not only contain the spatial forms but also the visualinformation based on the snapshots. As a result of the entire reconstruction a virtual object is gained in the virtual space that can freely be accessed and visible from all directions by enlarging or reducing the size. A virtual collection can be established for monuments, historical buildings or other spectacular objects worth recording. The objects of the collection (buildings) are lifelike and can be perceived and studied by anyone. The 3D models, of course, cannot substitute the photographs of high resolution but they can complement them in the collection.

1. Introduction

The fundamentals of photogrammetry have been in existence since 1858 [1]. It is interesting to note, however, that surveys based on photographs were first carried out for buildings. Surface models in aerial photography were given birth only later. The procedure of photogrammetry basically calling for a lot of calculations has been made widely accessible to the users due to computers of high performance and GPGPU cards that make calculations significantly faster. A kind of revolution is also noticeable with regards to aerial photography due to the spread of the cheaper and cheaper remote controlled flying objects of higher standard that carry cameras. These two technological developments made aerial photography and processing photos by computers widespread.

The aerial photographs taken by unmanned aerial vehicles (UAV) can substitute for the traditional aerial photography by airplanes. Following the processing of pictures the resolution of the photo fitted together can be 3-3.5 cm/pixel (10-14 Mpixel native resolution) even by using standard cameras. While processing overlapping photographs [2] (overlapping of 60% is required alongside all edges) the 3D relief map of the area can also be made in addition to the most frequently used orthophotos. According to experience in measuring it can be reduced even below 2-3 cm by using multi-rotor devices when taking an up-close picture (3-5 m) of a smaller object (building, monument etc.) specially. The results of 3D surveys can be used in several areas of the recording systems including archiving the temporary state of certain formations, establishing virtual museum for buildings, monuments and relics of architecture, monitoring and recording the processes of real estate investments, making records of public utilities, registering the output of surface mines, surveying and recording the recultivation stage of abandoned surface mines etc [3,4,5].

The paper is focused on the possibilities of making 3D pictures of different architectural monuments and their presentation in a virtual exhibition by presenting real surveys [6].

2. Photogrammetry

Photogrammetry is almost as old as photography. In 1858 when photochemical picture recording became well-known [7] Albrecht Meydenbauer a young architect resulting from his fortunate accident worked out a procedure where measuring and surveying buildings became possible on the basis of photos taken

*Andras Molnar, , Email: molnar@uni-obuda.hu

of them. The term photogrammetry is also linked to Meydenbauer. At first the term was published in an anonymous professional article of Wochenblatt des Architektenvereins zu Berlin in 1867 [8]. It was only later disclosed that the author of the article was Meydenbauer. The procedure is based on elementary geometrics by making use of the linear spread of light and relating the photo to the object being photographed by taking the camera and its optical parameters into consideration. Imaging itself is not too complicated but in the case of complex objects, calculations on its corner points are extremely time-consuming. It is worth noting that originally Meydenbauer worked out photogrammetry for surveying buildings but later on this method became widespread by processing aerial and space photographs. By using this method planar and orthophotos can be taken which have real significance in photography. Basically, orthophotos mean the aerial pictures taken in photography on which factual measures (distance, area) can be made. In many cases these methods help make maps designed by traditional methods more precise. The photogrammetry procedure for the 3D model construction of buildings has become more popular with the spread of computers of powerful calculating performance. The reason lies in the fact that fitting a lot of parts of photographs together manually is demanding but the use of traditional computers was very restricted (in the case of few photos of small resolution) for automation. In general, good quality textured 3D models require the analysis and processing more than ten Gigabytes data.

3. Making 3D models on photogrammetric basis

The necessary prerequisite of making photogrammetric 3D models [9] is to have several photographs of the same relief from slightly different positions. This can happen when such aerial pictures are taken in practice where overlapping between frames is ensured with all the imaged pixels appearing at least in two pictures (when special boundary conditions are met). However, in practice pixels must appear in more than two photos. If the spatial position of recording the image is known together with the direction vector of the optical axis of the camera, the line running through the projection plane and defining the pixel concerned can be described. If this pixel can also be found in another picture, then another line also runs through the projection plane but its point of intersection is not identical with the one of the projection plane of the previous pixel disregarding some special cases. It is made possible by the fact that the same pixel observed from different perspectives is seemingly placed elsewhere. At the same time, the two projection intersecting lines define a point outside the projection plane that corresponds with the spatial position of the pixel in question.

Two basic requirements of defining projection beams are knowing the exact position of recording the image and the optical axis of the camera together with the pixel pairs (the same pixels) of the overlapping pictures. The position of recording the image can derive from the sensor system built in the camera or attached to it. Basically, this sensor system means 3D coordinates of summa rising GPS (Global Positioning System) and IMU (Inertial Measurement Unit) data and a unit for orientation. However, these data can also be defined by the visual information of the images recorded. With regard to the fact that due to the reasons listed above the pictures are taken with significant necessary overlapping the optical axis of the camera and its relative spatial position can also be defined based on their content analysis. Provided that the central pixel of the image from a geometrical aspect is free from distortion in each picture and it is regarded as the reference point between the pictures, special images in a different position can also be found in several pictures. The distance of these pixels from the reference point does not change in real although it differs from picture to picture. This makes the definition of such a transformation possible which gives the spatial position of recording the image. The method can be well automated but calls for a visual content accessible where the special pixels described above can be defined properly. Some examples include the corner points of buildings and intersections of roads or other objects discernibly separated from the background etc. The method cannot be used in homogeneous pictures such as riffle and water surfaces free from structured reflected images. From a practical point of view when examining surface objects in most cases the well-arranged and contrastive structure serve with enough reference points, fortunately.

The similar point pairs ensure connection between the images. Finding these point pairs can also be automated similarly to finding the position. The basis of searching for these point pairs is to find the so-called corner points in the pictures that stand apart from the other pixels. They are typically pixels on the border of strong intensity changes such as corners of buildings, borders of forests, highways etc. Certain characteristics can be assigned to these points such as intensity gradient and/or the intensity spread of adjacent pixels. Several corner point detecting algorithms can be used for that purpose including the most frequently used Harris-algorithm [10,11]. If these characteristics are invariant to any magnification or rotation, there is a possibility for finding more corner points with similar characteristics in the overlapping point pair. Similarly to detecting corner points, several algorithms are also available for defining such point pairs including SSD (Sum of Squared Difference) algorithm. The corner points founded and arranged in pairs must be further refined by filtering algorithms as at this stage of processing there are several faulty point pairs in the system. One of the best known filtering methods is RANSAC (RANdom SAmple Consensus).

By using the algorithms above a spatial point cloud can be created form the overlapping images that correspond with the discrete set of points of the surface of the surveyed area. The contiguous surface model is created by connecting the points together. Of course, while creating the surface model several algorithms can also be applied. One of the best known is called the Delaunay triangulation. As a result, a surface model covered by contiguous triangles can be derived. Special morphological filters containing the presumptions typical of the relief conditions can also be applied. These filters refine the strikingly high elements of the point cloud.

If the position of recording the images was defined on the basis of the visual content and not GPS the surface model created is not capable of making quantitative measurements yet. To this end, geo-referencing the data set created is required. The condition of geo-referencing is that the original area should have geodesic reference points or reference points exactly measured and identified in the pictures prior to recording. A less exact method is to define the coordinates of the well-identified objects in the picture by using other databases such as Google Earth. A 3D object

placed in the world and measurable is created as a result of geo-referencing.

There are customized systems for processing overlapping images that ensure all the services listed above and in half-automated way they create the 3D model of the surveyed area. Such systems include Agisoft PhotoScan [12].

4. The practical aspects of surveying buildings

The method applied when surveying areas for a general orthophoto is not suitable for surveying buildings. However, there is a surface model from the pictures of the same flight line which includes the extensions of the building by large in real but some details are never entered into the system. In the case of buildings there are several parts which are covered when taking an aerial photo.

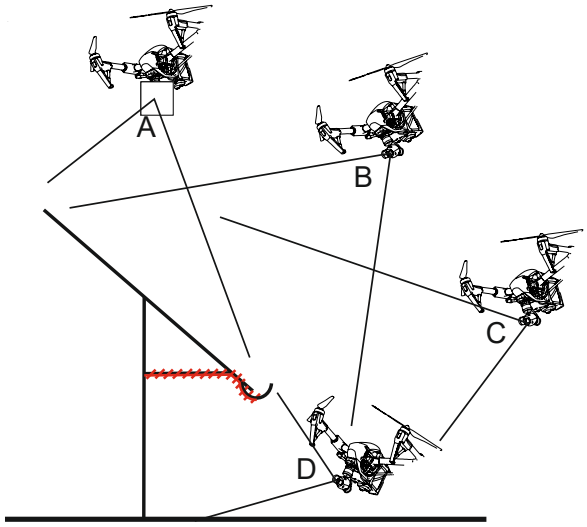


Figure 1: The visibility of the surface from different camera positions.

Figure 1 presents the visibility of surfaces when taking pictures of a typical building. It can generally be stated that taking photos from positions 'A', 'B' and 'C' does not mean a problem. At the same time, however, it can be seen that from these positions the bottom of the eaves marked by a stripe can never be seen so a real surface of this area cannot be taken in 3D reconstruction. Camera position 'D' is necessary to make pictures of the nonvisible parts. Unfortunately, in several cases this position is risky as the multirotor device carrying the camera stays very close to both the building and the ground. In the case of high buildings position 'D' is relatively easy to carry out as the minimum height of the multirotor above the area is higher, too. If it is feasible that the flying device should not go lower than 1-2 metres, pictures from position 'D' can also be taken. In the case of lower buildings taking pictures from position 'D' is also feasible if the drone is taken by hand and setting its camera to a proper direction photos of the building can be taken.

In the case of buildings the lateral and, in many cases, upward camera position is of high importance. That is why the fixed-wing aircrafts are not suitable for this task. Multirotor devices are the most efficient camera carriers at the current level of technological development when surveying buildings. Although helicopter-like devices are able to perform a similar movement to that of multirotor devices but their mechanical design is much more complex. They are more prone to damage so their transportationis more complicated. The only disadvantage of multirotor devices is that they are unable to auto-rotate. It means when the power batteries are out these devices cannot fly. As flight is very close to buildings when surveying only the most developed multirotor devices with the safest features should be used. As a result of most recent developments the control panels of the 8-rotor-multicopters are able to stand in for an engine while properly controlling the other working ones. Using such instruments can significantlyimprove the safety of surveys. As the primary position data of multirotor flying objects are processed on the basis of on-board GPS good reception is essential. Unfortunately, it is the low flight and the covering of the building that can negatively influence GPS reception. Weak GPS data result in the multicopter's apparent swift and imbalance, which has to be compensated by the operator. Although this position keeping flaw is not dangerous on its own, it can really be a cause for concern due to the proximity of the building or any other objects. During flights close to objects it is not practical to apply automated flight linesas its precision is not satisfactory. In many cases blocked live images of the camera must also be taken into consideration as higher buildings can shadow the frequency of the video signal carrier. For security reasons it is practical to plan flights during which the operator can see the multicopter. In practice, it means that the flying object must be accompanied round and round the object on the ground.

The automated point pair search in the pictures makes standard photo taking from all directions unnecessary. Attention must only be paid that every part of the given building should be recorded in as many pictures as possible so a lot of overlapping pictures should be taken.

Practical experience says that pictures of 3-4 m are necessary for good quality reconstruction by means of a 12 megapixel camera free from distortion.

Figure 2 illustrates the process of photographing a small chapel. The positions of making the single pictures can be seen above the building in the picture together with the optical axis of the camera at the moment of taking pictures. It can also be seen that in the survey the pictures that should have been taken upward next to the building from position 'D' of Figure 1 are missing. Despite the missing photographs 3D reconstruction could be carried out although at some parts of the building it is of bad quality (Figure 3).

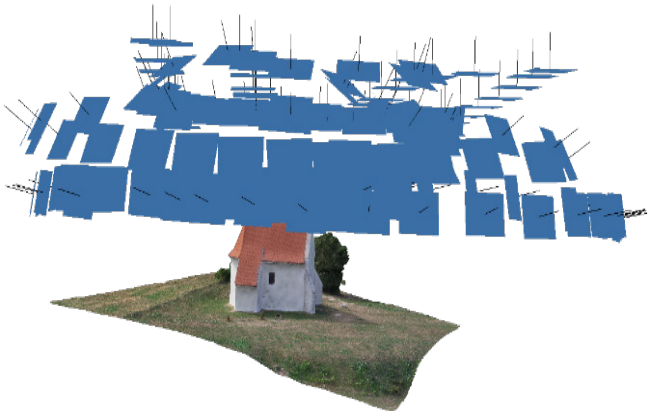


Figure 2: The spatial position and orientation of pictures taken when surveying a chapel.

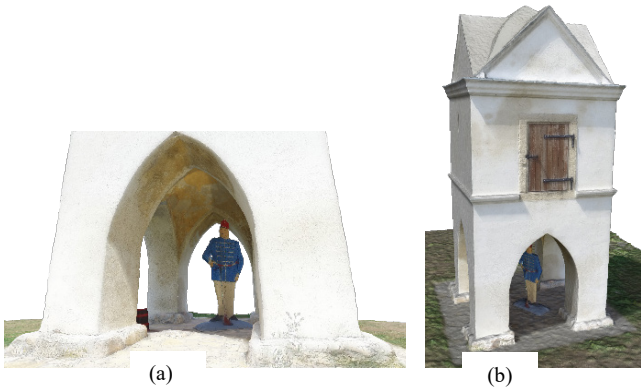
The order of making photos is not fixed, which means that pictures can be taken discretionary even randomly. The only criterion is that all parts of the object must be photographed and each part should appear in several pictures. In the case of a bigger building this requirement can be met if pictures are taken systematically such as taking pictures around the building (Figure 2). The practical implementation of going round the building can sometimes be defined by the nature of the building in many cases. The tower in Figure 4 is a narrow but high building. By choosing the more favourable and simpler way when photographing, pictures are taken 'up and down' alongside the four sides. It is worth noting that while taking pictures of the tower (Figure 4) the photos taken from position 'A' of Figure 1, i.e. above the top of the tower, are missing. However, 3D construction was successful as the pictures taken of the entire roof structure from positions 'B-C' of Figure 1 have the necessary overlapping.

Figure 3: Part of the reconstructed chapel under the eaves full of serious flaws.

In an extreme case pictures randomly taken are also suitable for 3D reconstruction provided the meet the requirements outlined above. At the same time, in the case of random photographs part of an object may sometimes be missing or not apparent in several pictures so that is why systematic photography is recommended.

Figure 4: The process of taking pictures of a high tower.

Figure 5 presents the 3D model of such a building where parts nonvisible from above are significantly represented. As a result of applying position 'D' of Figure 1 the vaulted joist of the passage of the building is also visible (Figure 5 (a)). Of course, this part of photography calls for utmost care as flights should be carried out extremely low and close to the building. In cases if the suspension of the camera system of the robot aircraft makes it possible, pictures can be taken by the manually rotating the multicopter (as if it was a handicam). It is important that the pictures taken 'manually' should be taken by a similar camera system and image fittings must be flawless.



(a) (b)

Figure 5: Display of parts of the historical tower non-visible from above.

The building to be photographed cannot always be freely accessible. If the walls of the building are at least laterally visible it is possible to make a correction of limitedly removing the trees or other covering object while processing.

Figure 6: The tree next to the wall of the building impedes photography.

The tree on Figure 6 does not make it possible for the multicopter to make a picture of the wall of the building freely. At the same time, however, the tree doesnot over the wall laterally so part of the wall covered from the front can be constructed from the necessary number of lateral photos. If no correction is made in processing,the 3D model will contain the tree and the wall, as well. By making use of the fact that the entire wall can be reconstructed the point cloud making up the tree can be removed during the 3D edition. It is important that such corrections be made on the point cloud serving as the basis of the model and not on the model ready. The 3D model of Figure 7 was made after the digital removal of the tree. The shadow of the tree is visible on the model but the tree itself no longer covers the building, so it is accessible.

Figure 7: The result of digitaly removing the tree in the 3D model.

The spatial point cloud generated while processing is connected with elementary triangles by the programme. These triangles make up the surface, which is the 3D model of the area already surveyed. Of course, the number of polygonsor any other means to detail the surface canbe stated here. The ready-made 3D model can be displayed from a discretionary point. The model contains only surface, i.e. the polygons covering the surface have no thickness. In a further step of processing texture can be fitted on the 3D surface. The texture is made by the programme by means of the photo analysed in the previous steps. If we take the chance of correcting the 3D point cloud then the automated use of the pictures serving as the basis of the texture is practical to be manually modified. In this concrete case removing the tree itself is not enough. To have the proper texture the parts depicting the tree must be excluded from the photos of the given area. In this case the program prepares the texture of the part concerned based on the pictures of the wall if they exist.

Figure 8: Ruins of a church from the Árpád era following retouching the neighboring trees.1

The 3D surface model (Fiure 8) is created by exporting the ready-made 3D model and one or more image files containing the texture (it depends on the size and the details of the model itself). All these files are necessary to display the D objects presented. Display itself is possibly by using several programmes such as the freely accessible and downloadable MeshLab programme.

Figure 9: The mausoleum of the Luppa family in Pomáz.2

5. Further3D models

Figure 9 presents the 3D reconstructed pjoto of the mausoleum of the Luppa family from Pomáz. The model of the mausoleum was made on the basis of 160 photos of 12 megapixels by using Agisoft Photscan software. The pictures were taken by a DJI Inspire 1 drone within one take-off of 15 minutes of flight. By walking round it in the virtual space the demolished state of the mausoleum is visible. From an aesthetic point of view several graffiti ruin the sight of the monument, which are also displayed on the reconstructed photo. It is interesting to note that the graffiti mentioned above are advantageous from the point of view of 3D reconstruction as it is easier to find contrasting point pairs when fitting the parts of the picture together than in the case of homogenous whitewashed walls.

Figure 10: A deserted farmhouse near Nádudvar.3

Figure 10 presents the 3D reconstructed picure of a deserted farmhouse. The farmhouse is approximately 8,2 m high, 25,8 m long and 10,8 m width. The model of the house is based on 160 photos of 12 megapixels by using Agisoft Photscan software. The pictures were taken by a DJI Inspire 1 drone within one take-off of 15 minutes of flight. When these photos were made, it was gloomy so the contrast of the reconstructed picture is a bit beyond the ones made in sunny weather but it proved to be advantageous for

[1]Retrieved from https://sketchfab.com/models/8dda3b73b7a14c2cb606b15ef75408c1
[2]Retrieved from https://sketchfab.com/models/6ee2ae74e48846c6a26e2884a51ac9c5
[3]Retrieved from https://sketchfab.com/models/5de27c3c8ad442d7bb4122fafcd268b6

reconstruction as there were no shadows and lighting proved to be constant when going round the entire building. The model reconstructed is detailed enough to reflect the state of the house. The mortar coming off and the bricks used for the walls staying bare are visible.

Figure 11: A renovated windmill in Kulcs.[4]

Figure 11 presents the 3D reconstructed photo of a renovated windmill. The model of the windmill is based on 160 photos of 12 megapixels by using Agisoft Photscan software. The pictures were taken by a DJI Inspire 1 drone within one take-off of 15 minutes of flight. hen taking pictures, utmost care must be paid to the thin wooden structure of the mill as reconstructing thin parts is usually difficult.

Figure 12: Papd chapel.[5]

Figure 12 presents the 3D reconstructed picture of a dilapidated chapel. The chapel is approximately 15,7 m high, 13,3 m long and 7,3 m width. The model of the chapel is based on 285 photos of 12 megapixels by using Agisoft Photscan software. The pictures were taken by a DJI Inspire 1 drone within two take-offs of nearly 30 minutes of flight. When the pictures were taken it was typically sunny but the sun was shaded by clouds for a short time. Due to this fact, both sunny and cloudy pictures were taken of the building, which proved to be useful in 3D reconstruction. As a result, the ready-made model is contrastive but free from the distortions of heavy shadowing. Reconstructing the cross at the top of the church requires special care and reconstruction of proper quality needs up-close photos of the cross and a small part of the dome. To this end, pictures of the top of the tower were made while the multicopter went round the cross at a distance of 4 metres. Pictures were once taken from the bottom and from the top of the cross. Despite the up-close photos the manual filtering in the 3D point cloud was necessary during the reconstruction of the cross due to which the points created as faults by the programme were removed.

6. Results

Based on the experience of several experiments the requirements of successful 3D reconstruction can be outlined.

Regarding environmental conditions diffuse light is thebest for reconstruction, which corresponds with slightly cloudy weather. Suitable pictures in direct sunlight can only be made if the object moderately reflects light of a non-homogenous, i.e. well-structured surface. The 3D reconstruction of the shadowy side of objects with homogenous white surface such as several churches and chapels could only be carried out in significantly weaker quality (roughness of surface) in direct sunlight.

Regarding the cameras applied the parameter of the objectives is of great significance in addition to high resolution (of at least 12 Megapixels). Most pictures were taken by an objective of 20 mm working distance of, 12.4 Megapixels with 6.17x4.55 mm CCD sensor. In many cases very proximate fights were necessary to provide enough details. It can be proved by practice that experiments yield better results with an objective of 34 m working distance of 16.0 Megapixels with 17.3x13.0 mm CCD sensor. In case of the latter one there is no need for dangerously approaching the object to have the necessary details.

Follow-up filtering proved to be necessary for several objects. In all cases filtering meant removing excess or disturbing points of the 3D point cloud. In some cases parts of the pictures deleted had to be removed, as well. However, it is important to note that the preliminary image improvement of the pictures used for the reconstruction significantly damaged the final result of the reconstruction. That is why exclusively unprocessed pictures could be used for processing.

The 3D models of the paper and further models of drone pictures are accessible at https://sketchfab.com/fuhur.

References

[1] J. Albertz, Albrecht Meydenbauer - Pioneer of Photogrammetric Documentation of the Cultural Heritage, Proceedings 18th International Symposium CIPA 2001, pp. 19-25, Potsdam (2001).

[2] M. Marghany, M. R. B. M. Tahar, M. Hashim, "3D stereo reconstruction using sum square of difference matching algorithm", Scientific Research and Essays, Vol. 6(30), 6404-6423 (2011) [doi:10.5897/SRE11.1661].

[3] Ildiko Horvath, Anna Sudar, „Factors Contributing to the Enhanced Performance of the MaxWhere 3D VR Platform in the Distribution of Digital Information", Acta Polytechnica Hungarica Vol. 15, No. 3, 2018 [doi: 10.12700/APH.15.3.2018.3.9].

[4] Borbála Berki, 2D Advertising in 3D Virtual Spaces", Acta Polytechnica Hungarica Vol. 15, No. 3, 2018doi: 10.12700/APH.15.3.2018.3.10].

[4]Retrieved from https://sketchfab.com/models/ed87fd62fc1b4e1a90293192355c8a65

[5]Retrieved from https://sketchfab.com/models/5ea66604b8a2462ebc499887c9fc2f08

[5] Şasi, A , Yakar, M., "Photogrammetric Modelling of Hasbey Dar'ülhuffaz (Masjid) Using An Unmanned Aerial Vehicle", International Journal of Engineering and Geosciences, 3 (1), 6-11 (2018) DOI: 10.26833/ijeg.328919
[6] Doğan, Y, Yakar, M., "Gis and Three-Dimensional Modeling for Cultural Heritages". International Journal of Engineering and Geosciences, 3 (2), 50-55 (2018). DOI: 10.26833/ijeg.378257
[7] H. Gernsheim, "The 150th Anniversary of Photography," in History of Photography, Vol. I, No. 1, (1977).
[8] A. Grimm, "Der Ursprung des Wortes Photogrammetrie", in Internationales Archiv für Photogrammetrie, Ackermann, F. et al., Ed., Vol. XXIII, Teil B10, Komm. V, VI, VII Nachtrag, pp. 323-330, Hamburg (1980).
[9] N. A. Matthews, Aerial and Close-Range Photogrammetric Technology: Providing Resource Documentation, Interpretation, and Preservation. Technical Note 428. U.S. Department of the Interior, Bureau of Land Management, National Operations Center, 42 pp., Denver, Colorado (2008).
[10] C. Harris, M. Stephens, "A combined corner and edge detector", in 4th Alvey Vision Conference, Proc. AVC88, 147-151 (1988) [doi:10.5244/C.2.23].
[11] A. Willis, Y. Sui, An Algebraic Model for fast Corner Detection, in 12th International Conference on Computer Vision, Proc. ICCV.2009, 2296-2302 (2009) [doi:10.1109/ICCV.2009.5459443].
[12] Ulvi, A , Toprak, A., "Investigation of Three-Dimensional Modelling Availability Taken Photograph of The Unmanned Aerial Vehicle; Sample Of Kanlidivane Church", International Journal of Engineering and Geosciences, 1 (1), 1-7 (2016) DOI: 10.26833/ijeg.285216

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 258-267 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Performance Investigation of Semiconductor Devices using Commutation-speed based methodology for the application of Boost Power Factor Correction

Barkha Parkash*,[1,2], Ajay Poonjal Pai[1], Wei Tian[2], Ralph Kennel[2]

[1]*Infineon Technologies AG, Automotive High Power, 85579, Germany,*

[2]*Technical University of Munich, Chair of Electrical Drive Systems and Power Electronics, 80333, Germany*

ARTICLE INFO

Article history:
Received: 06 December, 2018
Accepted: 03 February, 2019
Online : 20 February, 2019

Keywords:
Commutation-speed
Power losses
Switching losses
Analytical model

ABSTRACT

In this paper, behavioral approach has been adopted for the calculation of total power losses that has been further used to derive an analytical model for the conduction and switching losses in a boost Power Factor Correction (PFC) stage of an On-board Charger (OBC). Detailed investigation of power losses can help in finding out ways to improve efficiency and for this purpose, commutation-speed based methodology has been used to split total power losses into their root causes. This gives opportunity to find the impact that an individual part creates on total losses which can serve as a starting point for efficiency improvement. For the analysis, two devices (IGBT with Si diode and IGBT with SiC diode) are used in the considered topology of PFC and a reduction of 40% was calculated when SiC diode was used instead of Si with the same IGBT. Hence it was found that the implemented method proves to be significantly useful in the optimization of efficiency.

1. Introduction

In electric vehicles (xEVs), it is desirable for the on-board charger to operate under maximum efficiency. Therefore, it is of prime importance to accurately calculate the power losses at the system level. A significant part of the power losses arise from the semiconductor devices, being used. A number of different methods are available for the calculation of power losses in semiconductor devices. One of the methods available is physics-based, which requires precise models of the power semiconductor devices and circuits under consideration are simulated numerically with the help of specially designed programs [1]. However in this method, the simulation time tend to be very long which doesn't seem to be a practical option for system simulations.. Another widely used approach which has been adopted in this paper, is the behavioral modelling of the power losses in which behaviour of the semiconductor device is captured under different operating conditions which is further used to develop simple equations to model the losses. This approach has been proven to give accurate results as discussed in [2,3]. It not only supports simplified calculation but also allows the application of commutation-speed based method in which all devices are switching at the same speed. This methodology makes it possible to split total power losses into their root causes, e.g. capacitive effect, tail current, reverse recovery, forward conduction, reverse conduction [4]. The accurate split-up of power losses give the opportunity to assess and optimize the switching behaviour of power semiconductors at the converter level.

*Corresponding Author: Barkha Parkash, barkhaparkash2@gmail.com

2. Loss Modelling

Power loss modelling involves creation of a generic model that helps in the accurate calculation of total power losses. The main losses associated with a semiconductor device are conduction, switching and gate driving losses but gate driving losses are not discussed in this paper. In general, the loss modelling is done using behavioural model of the device. For the behavioural part, several static and dynamic characterization measurements are conducted that gives insight into the behaviour of the device under different operating conditions. Characterization measurements are performed on the double pulse setup as discussed in [5]. The used setup is shown in Figure 1. For the characterization, one parameter is varied at a time while all other parameters are kept constant at their nominal values. For a higher degree of accuracy, the power loss equations are derived using a model that calculates losses as a function of device current $I_{c/f}$, DC link voltage V_{dc}, junction temperature T_j, gate resistance R_g, area of device $A_{d/l}$ and gate voltage V_{ge} as elaborated in [2]. The derived equations are used for the calculation of power losses in a boost Power Factor Correction (PFC) stage of an On-board Charger (OBC).

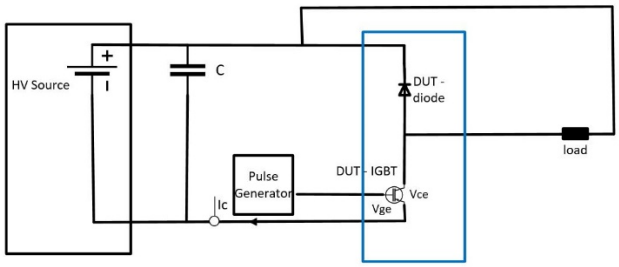


Figure 1 Double pulse setup for device characterization

2.1. Conduction Losses

Conduction losses are determined using instantaneous current and corresponding forward voltage. The simplest and most widely used way to model forward voltage is first order linear approximation consisting of threshold voltage and the drop across series resistance [6]. However, this model is accurate at the nominal operating point whereas the accuracy reduces at all other non-nominal points. A better approach to model forward voltage, which is also used in this work, is to consider the dependency between the considered parameters and voltage as quadratic functions instead of linear. The main advantage of using this approach is its accuracy at non-nominal points as well.

Mathematically, it can be written as [2];

$$V_{ce} = f(J_c).\, f(T_j).\, f(V_{ge}) \tag{1}$$

Where,

V_{ce} = Collector to emitter voltage in the case of an IGBT

J_c = Current density in the chip

$$f(J_c) = A_{11}.J_c^2 + A_{12}.J_c + A_{13} \tag{2}$$

$$f(T_j) = A_{21}.T_j^2 + A_{22}.T_j + A_{23} \tag{3}$$

$$f(V_{ge}) = A_{31}.V_{ge}^2 + A_{32}.V_{ge} + A_{33} \tag{4}$$

Similarly, the forward voltage of diode V_f is modelled as;

$$V_f = f(J_d).\, f(T_j) \tag{5}$$

Where,

J_d = Current density of the diode

2.2. Switching Losses

The most accurate way to calculate switching losses is to integrate the area where both voltage and current are overlapping in transition region, but this requires detailed knowledge of transient curves, which isn't readily available in most cases. A work around of this approach is to use the commutation time in which device is turned on or off and to take the corresponding voltage and currents. This approach is comparatively easier because this information is given in the datasheet [2]. However, these models are fitted at the nominal point and there is increasing error at all other points. That's why, this paper proposes to use the model presented in [2] where the switching energies as a product of quadratic functions of the considered parameters.

$$E_{on} = f(J_c).f(T_j).f(V_{dc}).f(R_g).f(A_I).f(A_D) \tag{6}$$

Where,

$$f(J_c) = C_{11}.J_c^2 + C_{12}.J_c + C_{13} \tag{7}$$

$$f(T_j) = C_{21}.T_j^2 + C_{22}.T_j + C_{23} \tag{8}$$

$$f(V_{dc}) = C_{31}.V_{dc}^2 + C_{32}.V_{dc} + C_{31} \tag{9}$$

$$f(R_g) = C_{41}.R_g^2 + C_{42}.R_g + C_{43} \tag{10}$$

$$f(A_I) = C_{51}.A_I^2 + C_{52}.A_I + C_{53} \tag{11}$$

$$f(A_D) = C_{61}.A_D^2 + C_{62}.A_D + C_{63} \tag{12}$$

Similarly, turn off and reverse recovery energy are given as following;

$$E_{off} = f(J_c).f(T_j).f(V_{dc}).f(R_g).f(A_I) \tag{13}$$

$$E_{rec} = f(J_f).f(T_j).f(V_{dc}).f(A_D) \tag{14}$$

2.3. Extraction of polynomials for the behavioral model of device

These coefficients (A_{11},A_{12} ,A_{13}…C_{63}) are determined by fitting a quadratic curve to the characterization measurements. They represent the behaviour of the device and are used to determine the switching and conduction losses at any operating point.

3. Analytical Model for the losses in Boost PFC

Figure 2 shows boost PFC, which is a commonly used PFC topology in an on-board charger. In a boost PFC, there are two sub-sections; rectifier followed by the boost stage that is controlled to ensure unity power factor. Based on the discussion in section 2, an analytical model for the switching and conduction losses of the switching devices in a boost PFC are derived.

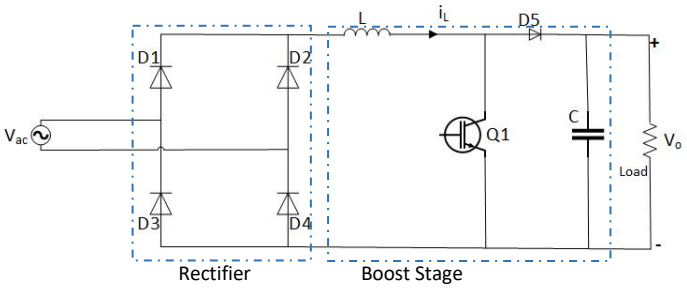


Figure 2 Boost PFC stage of an On-board Charger

The inductor current is sinusoidal; therefore, it can be written as a function of time.

$$i_L = I_{pk}.\sin\alpha \tag{15}$$

Where I_{pk} is the peak value of current and α is given as;

$$\alpha = w.t \tag{16}$$

The duty cycle for the switch $\delta(t)$ in boost PFC is given by;

$$\delta(t) = 1 - \frac{V_{pk}\sin\alpha}{V_o} \tag{17}$$

V_{pk} is the peak value of input voltage and V_o is the output voltage which is equal to the V_{dc}. Whereas the duty cycle for the boost diode $\delta_{diode}(t)$ is given by;

$$\delta_{diode}(t) = \frac{V_{pk} \sin \alpha}{V_o} \quad (18)$$

So, the instantaneous switch i_Q and diode current i_{diode} are given as;

$$i_Q = I_{pk}.\sin \alpha. \{1 - (\frac{V_{pk}\sin \alpha}{V_o})\} \quad (19)$$

$$i_{diode} = I_{pk}.\sin \alpha. (\frac{V_{pk}\sin \alpha}{V_o}) \quad (20)$$

The instantaneous conduction losses for the switch and diode in boost PFC are as follows;

$$p_{s_{cond}} = V_{ce} \cdot i_Q \quad (21)$$

$$p_{s_{diode}} = V_f \cdot i_{diode} \quad (22)$$

where,

$$V_{ce} = (A_{11}.J_c^2 + A_{12}.J_c + A_{13}). \left(A_{21}.T_j^2 + A_{22}.T_j + A_{23} \right).\left(A_{31}.V_{ge}^2 + A_{32}.V_{ge} + A_{33} \right) \quad (23)$$

$$V_f = (B_{11}.J_f^2 + B_{12}.J_f + B_{13}).\left(B_{21}.T_d^2 + B_{22}.T_d + B_{23} \right) \quad (24)$$

The average conduction loss in the switch $P_{s_{cond}}$ is calculated as given in paper [7]. Integration is performed over half the fundamental cycle and due to symmetrical nature, the integration result is directly multiplied with 2 and then averaged over the fundamental cycle. Mathematically, it is given as;

$$P_{s_{cond}} = \frac{2}{T}\int_0^{\frac{T}{2}} p_{s_{cond}}\, dt \quad (25)$$

Where, T is the fundamental period. On substitution, following expression is obtained.

$$P_{s_{cond}} = \frac{2}{2\pi}\left\{\int_0^{\pi}((A_{11}.J_c^2 + A_{12}.J_c + A_{13}).\left(A_{21}.T_j^2 + A_{22}.T_j + A_{23} \right).\left(A_{31}.V_{ge}^2 + A_{32}.V_{ge} + A_{33} \right)) . I_{pk}.\sin \alpha \left(1 - \frac{V_{pk}\sin \alpha}{Vo}\right) d\alpha \right\} \quad (26)$$

Following result is obtained on performing integration.

$$P_{s_{cond}} = \frac{K_c.J_{cpk}.A_i}{\pi}\left[\frac{4.A_{11}.J_{cpk}^2}{3} + \frac{\pi.A_{12}.J_{cpk}}{2} + 2.A_{13} - \frac{3.\pi.A_{11}.J_{cpk}^2.V_{pk}}{8.V_o} - \frac{4.A_{12}.J_{cpk}.V_{pk}}{3.V_o} - \frac{A_{13}.V_{pk}.\pi}{2.V_o}\right] \quad (27)$$

Where, K_c is a constant which is given by;

$$K_c = \left(A_{21}.T_j^2 + A_{22}.T_j + A_{23} \right).\left(A_{31}.V_{ge}^2 + A_{32}.V_{ge} + A_{33} \right) \quad (28)$$

Similarly, the diode conduction losses are derived as;

$$P_{D_{cond}} = \frac{K_d.J_{d_{pk}}.A_d}{\pi}\left[\frac{3.\pi.B_{11}.J_{d_{pk}}^2.V_{pk}}{8.V_o} + \frac{4.B_{12}.J_{d_{pk}}.V_{pk}}{3.V_o} + \frac{B_{13}.V_{pk}.\pi}{2.V_o}\right] \quad (29)$$

Where, constant K_d is given as;

$$K_d = \left(B_{21}.T_d^2 + B_{22}.T_d + B_{23} \right) \quad (30)$$

The average switching losses is the sum of turn on, turn off and recovery loss. Integrating switching energies as given by (5), (12) and (13); following results are obtained.

$$P_{on} = f_{sw}\left[\left(\frac{C_{11}.K_{on}}{2}\right).J_{c_{pk}}^2 + (C_{12}.K_{on}).\frac{2.J_{c_{pk}}}{\pi} + (C_{13}.K_{on})\right] \quad (31)$$

$$P_{off} = f_{sw}\left[\left(\frac{D_{11}.K_{off}}{2}\right).J_{c_{pk}}^2 + (D_{12}.K_{off}).\frac{2.J_{c_{pk}}}{\pi} + (D_{13}.K_{off})\right] \quad (32)$$

$$P_{rec} = f_{sw}.K_{rec}\left[\frac{E_{11}.J_{d_{pk}}^2}{2} + \frac{2.E_{12}.J_{d_{pk}}}{\pi} + E_{13}\right] \quad (33)$$

Where, f_{sw} is the switching frequency and constants K_{on}, K_{off} and K_{rec} are given as;

$$K_{on} = \left(C_{21}.T_j^2 + C_{22}.T_j + C_{23}\right).\left(C_{31}.V_{dc}^2 + C_{32}.V_{dc} + C_{33}\right).\left(C_{41}R_g^2 + C_{42}R_g + C_{43}\right).\left(C_{51}.A_I^2 + C_{52}A_I + C_{53}\right).\left(C_{61}.A_D^2 + C_{62}.A_D + C_{63}\right) \quad (34)$$

$$K_{off} = \left(D_{21}.T_j^2 + D_{22}.T_j + D_{23}\right).\left(D_{31}.V_{dc}^2 + D_{32}.V_{dc} + D_{33}\right).\left(D_{41}R_g^2 + D_{42}R_g + D_{43}\right).\left(D_{51}.A_I^2 + D_{52}A_I + D_{53}\right) \quad (35)$$

$$K_{rec} = \left(E_{21}.T_j^2 + E_{22}.T_j + E_{23}\right).\left(E_{31}.V_{dc}^2 + E_{32}.V_{dc} + E_{33}\right).\left(E_{41}.A_D^2 + E_{42}.A_D + E_{43}\right) \quad (36)$$

4. Splitting of Switching Losses

Experimental measurements give a complete picture of the total losses in the device however it doesn't explain the contribution of various factors that make up the total losses. To get a deeper insight into the switching losses, the switching events (turn on and turn off) are divided into various regions as explained in [4].

4.1. Turn On

The turning on of the device is divided into two regions as shown in Figure 3; commutation region where the actual current commutation occurs from the diode to the active switch and the reverse recovery region which is influenced by the type of freewheeling diode (FWD), being used.

4.1.1 Region A

The split-up can be achieved by utilizing the device current and voltage waveforms obtained by measurements. Considering device current and voltage shown in Figure 3, the commutation region starts when the gate voltage goes higher than its threshold and the collector current just starts increasing which is marked as t_o. This region continues till the collector current has reached its steady state value, which is equal to the load current and this time instant is indicated in the Figure as t_A. The area from time t_o to t_A is taken as Region A. During the current rise, the collector to emitter voltage drops and this voltage drop is due to the stray inductance of the commutation path, L_σ and the rate of change of current $\frac{di_{ce}}{dt}$ as described in [8];

$$V_{ce} = V_{dc} - L_\sigma.\frac{di_{ce}}{dt} \quad (36)$$

4.1.2 Region B

The second region, reverse recovery part starts when the current through free-wheeling diode (FWD) reaches zero but the diode cannot take the entire voltage yet because of the presence of plasma [9]. Therefore, the current through the active switch continues to increase. The region B constitutes the time when collector current continues to rise due to plasma of FWD till the complete removal of excess carriers and falls subsequently to the steady state value i.e. at time instant t_B. The maximum value of collector current depends on the design of diode, gate driver that influences the rate of change of current, the junction temperature and the dc bus voltage [8]. The area from time t_A to t_B is the Region B.

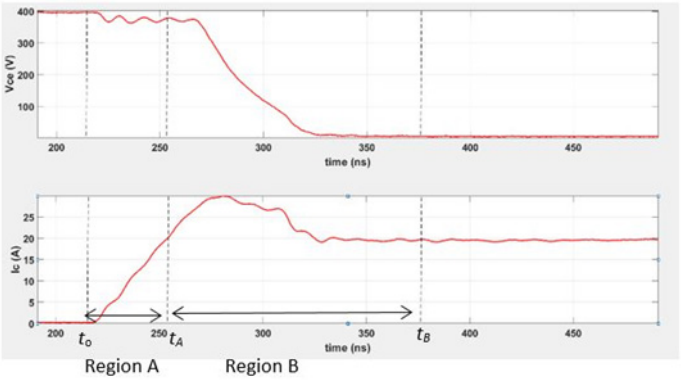


Figure 3 Split-up of the turn on event

4.2. Turn Off

The turn off event shown in Figure 4 is divided into three regions for an IGBT; capacitive region, commutation region and tail current and the speed at which turn off occurs depends on the gate resistance.

4.2.1 Region A

For the IGBT to turn off, it is important that diode takes up the load current which is not possible till voltage across diode has reduced to zero i.e. its junction capacitance needs to be discharged. The current that is used for the discharging of junction capacitance of the diode causes a gradual reduction in active switch's current i.e. collector current in the case of IGBT. This forms the capacitive region which can be seen in the Figure where current starts falling at time t_o and the active switch takes up the full voltage at time t_A.

4.2.2 Region B

After discharging of junction capacitance, the collector current starts falling rapidly at time t_A and that's when the actual commutation takes place from the active switch to the diode.

4.2.3 Region C

The bipolar devices like IGBT do not turn off till the minority charge carriers have been removed completely. The sweeping of minority carriers from the device takes some time resulting in tail current which starts at t_B and continues till current reaches zero, marked in the Figure as t_C. The tail current doesn't exist in the case of MOSFET because it is a unipolar device.

The different regions of turn on, turn off and diode switching losses add up to form total switching losses of the switch. If the contribution of each part on an individual level is known then it can help in the assessment of the impact that they create. This information can be very beneficial in optimizing the efficiency.

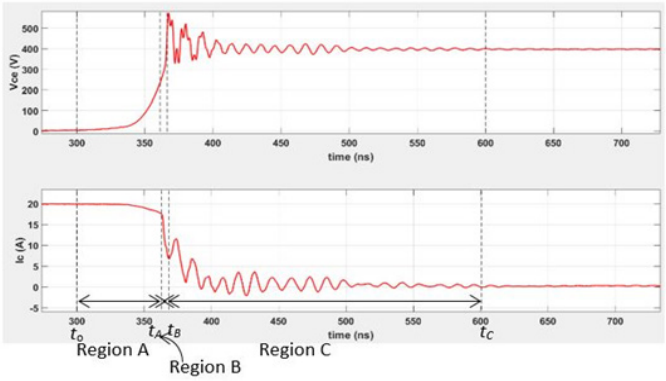


Figure 4 Split-up of the turn off event

5. Compared Devices

To understand the essence of this methodology, it has been implemented on two different discrete devices. One of the devices is Trenshtop-5 IGBT with Rapid Si diode from Infineon [10] and the second device is produced by replacing the Si diode of first device with SiC Schottky diode from Infineon whereas same IGBT is used for both the devices. In this paper, the first device will be referred to as "Full-Si IGBT" and the second device pair as "Hybrid-Si/SiC IGBT". The nominal values for both the devices are given in Table 1.

Table 1 Nominal values for both the considered devices

Nominal Values	I_c	T_j	V_{dc}	R_g
Full-Si IGBT	20 A	25°C	400 V	2.3 Ω
Hybrid-Si/SiC IGBT	20 A	25°C	400 V	14.6 Ω

6. Effect of Parameters on Split-wise Losses

All the parameters considered here; device current, junction temperature, gate resistance, DC link voltage, area of the device and gate voltage affect the switching losses. To get a better understanding as how each parameter effects the power losses in the devices, these devices are characterized using double pulse setup in which each of these parameter is varied at a time and corresponding total turn on and turn off energies are calculated. The data from characterization measurement is simulated on a specially written program that splits the total turn on and turn off energies into their respective parts as discussed in section 4. The split-up is done for both the considered devices. The results of the simulation are discussed as below where Full-Si IGBT is taken as the reference and the trends where Hybrid-Si/SiC IGBT differs are also explained. The comparison to study the effect of different parameters on switching energies isn't one to one because the nominal gate resistance is different for the devices.

6.1. Turn On

6.1.1 Effect of Device Current

It is observed that with increasing device currents, the total power losses due to turn on event increase. The trend of turn on energy as a function of device current for Full-Si IGBT and

Hybrid-Si/SiC IGBT are shown in Figure 5 and 6 respectively. It can be seen in these figures that the losses due to region A and region B are increasing with increasing current levels. As the value of device current is increased then the rate of change of current $\frac{di_{ce}}{dt}$ decreases i.e. current rises slowly as shown in Figure 7. For a current increase from 5A to 20A, the corresponding rate of change of current is from 0.65 kA/µs to 0.582 kA/µs. This indicates that the losses due to commutation (region A) increase with increasing current. The reason for the increase in losses is due to the direct relation between the value of current and the time taken by the device to reach the nominal value. For higher values of current, device takes more time to rise to that value.

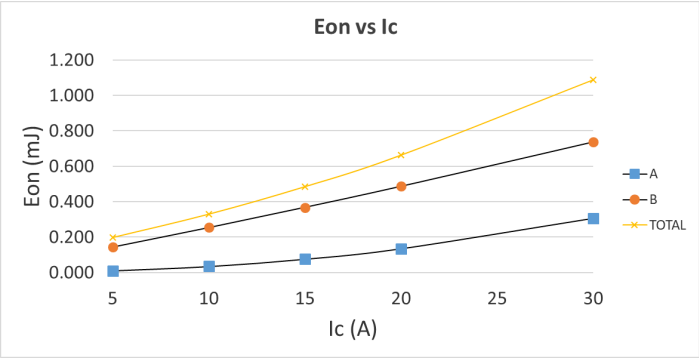


Figure 5 Turn on energy arising from its different root causes as a function of device current (Full-Si IGBT) at V_{dc} = 400 V, T_j = 25°C and R_g = 14.6 Ω A = commutation, B = reverse recovery

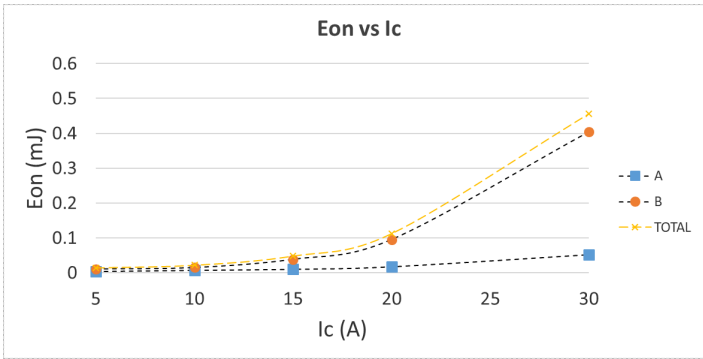


Figure 6 Turn on energy arising from its different root causes as a function of device current (Hybrid-Si/SiC IGBT) at V_{dc} = 400 V, T_j = 25°C and R_g = 2.3 Ω A = commutation, B = reverse recovery

Losses due to reverse recovery depends on the type of freewheeling diode. In the case of Si diode, which is a bipolar device, the losses due to this region are predominant because Si diode cannot support voltage till the removal of plasma from its PN junction. Therefore, the voltage at IGBT stays higher and allows collector current to rise beyond its nominal point which initiates the sweeping out of plasma from the junction [8]. Therefore, the losses due to reverse recovery (region B) increases with increase in current. Higher current means more number of charge carriers so the time taken by the device to sweep out the excess carriers from PN junction of freewheeling diode is greater as well which leads to increase in losses. On the other hand in the case of unipolar diode like Schottky SiC diode, the losses due to this region are very low because plasma doesn't exist here and the losses are due to the junction capacitance only.

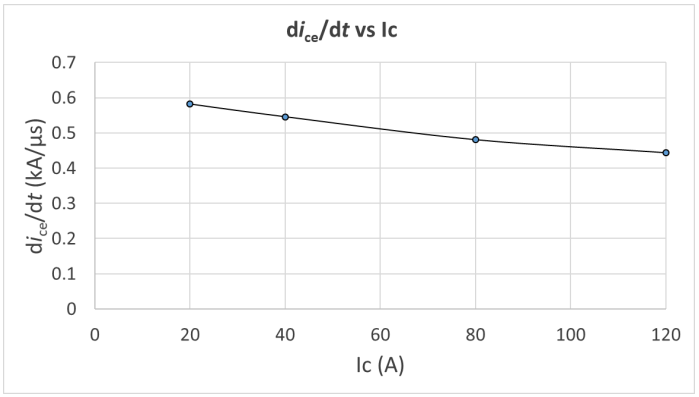


Figure 7 Rate of change of current as function of current (Full-Si IGBT) at V_{dc} = 400 V, T_j = 25°C and R_g = 14.6 Ω

It is to note here that in turn on losses, major contribution of losses is due to region B which corresponds to reverse recovery for Full-Si IGBT and charging of junction capacitance in the case of Hybrid-Si/SiC IGBT.

6.1.2 Effect of Junction Temperature

With increase in temperature, the threshold voltage at which the IGBT turns on decreases, which leads to faster commutation compared to lower temperatures. However, this has a small effect on the power losses. Increment in temperature leads to a slight decrement in power losses due to region A.

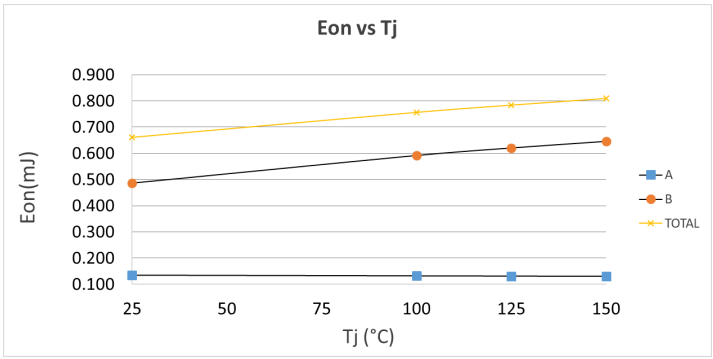


Figure 8 Turn on energy arising from its different root causes as a function of junction temperature (Full-Si IGBT) at V_{dc} = 400 V, I_c = 20 A and R_g = 14.6 Ω A = commutation, B = reverse recovery

As it can be seen in Figure 8, the losses due to region B increase with increase in temperature. This is because of the fact that the mobility of charge carriers decreases at higher temperature, which slows down the sweeping process of excess carriers from PN junction of diode, leading to increment in losses due to reverse recovery. So, it can be said that at higher temperature, major portion of turn on losses is due to the region B.

Figure 9 shows the dependency of temperature on turn on energy for Hybrid-Si/SiC IGBT. It is seen that the losses due to region A follow the same trend like Full-Si IGBT. This is because the region A is related with the rising current in the IGBT and same IGBT is being used for both the devices. However, the losses due to region B don't change with temperature. This is because in a SiC diode, the losses in this region do not come from reverse recovery but from the capacitive charge which is independent of the temperature.

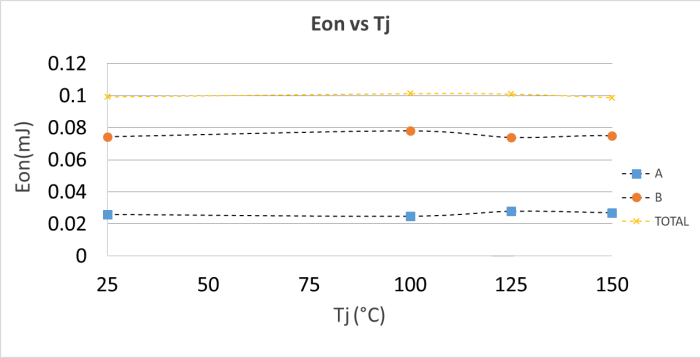


Figure 9 Turn on energy arising from its different root causes as a function of device current (Hybrid-Si/SiC IGBT) at V_{dc} = 400 V, I_c = 20 A and R_g = 14.6 Ω A = commutation, B = reverse recovery

6.1.3 Effect of DC link Voltage

During region A of turn on, the fall of collector to emitter voltage, V_{ce} is given by (36). The stray inductance L_σ is constant and $\frac{di_{ce}}{dt}$ doesn't change significantly at the nominal current and gate resistance. The only variable is V_{dc} which means that amplitude of V_{ce} increases with the increase in DC link voltage.

In general, the power losses are calculated as integration of the product of V_{ce} and i_{ce}. As V_{ce} is higher for higher values of V_{dc} so the power losses due to region A and B increase. However, the power losses due to region B increase by bigger margin compared to region A as shown in Figure 10. This happens because of the fact that the region B continues till the voltage has reduced to zero, so the time taken by the active switch for its voltage to drop to zero will be longer for higher values of V_{dc} .

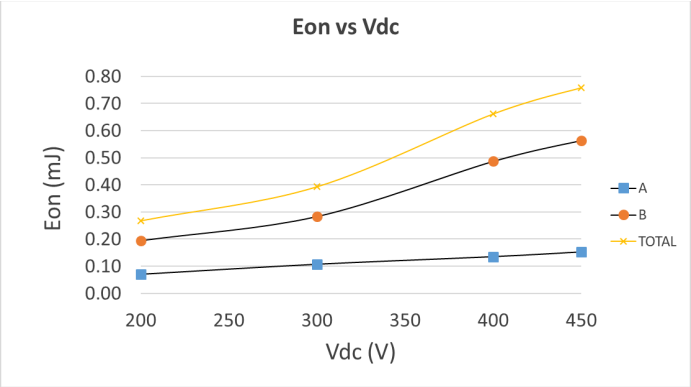


Figure 10 Turn on energy arising from its different root causes as a function of DC link voltage (Full-Si IGBT) at T_j = 25°C, I_c = 20 A and R_g = 14.6 Ω A = commutation, B = reverse recovery

6.1.4 Effect of Gate Resistance

The charging of the gate capacitance determines the switching behavior of the semiconductor device and this is controlled by an external gate resistance. The gate resistance has influence on switching losses; therefore, optimal value of gate resistance should be selected according to the application [11].

A smaller gate resistance leads to shorter rise time which means device switches faster. So, a big gate resistance leads to increment in current rise time i.e. rate of change of current reduces as indicated in Figure 11. This leads to increment in losses due to region A as shown in Figure 12.

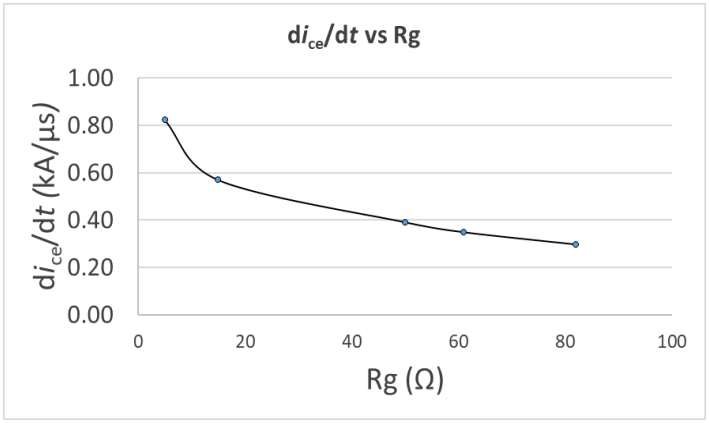


Figure 11 Variation in rate of change of current for different gate resistances at Tj = 25°C, Ic = 20 A and Rg = 14.6 Ω

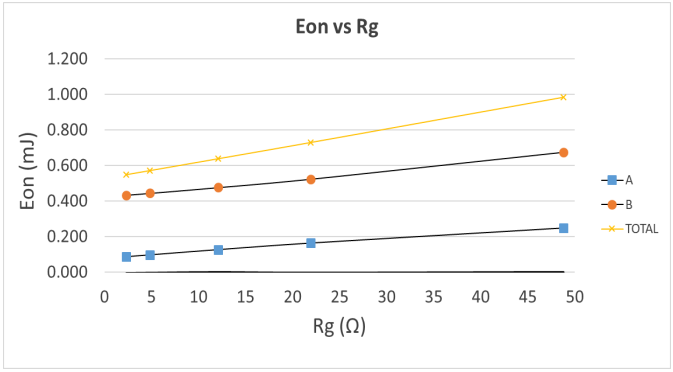


Figure 12 Turn on energy arising from its different root causes as a function of gate resistance (Full-Si IGBT) at V_{dc} = 400 V, I_c = 20 A and T_j = 25°C A = commutation, B = reverse recovery

With increasing gate resistance, there are two effects as shown in Figure 13; one is that the reverse recovery peak decreases and secondly $\frac{dV_{ce}}{dt}$ decreases which indicates that the time taken by the active switch for its voltage to drop to zero is longer compared to a smaller gate resistance. Out of these two effects, the predominant effect is the reduction in $\frac{dV_{ce}}{dt}$ which causes increment in the losses due to region B.

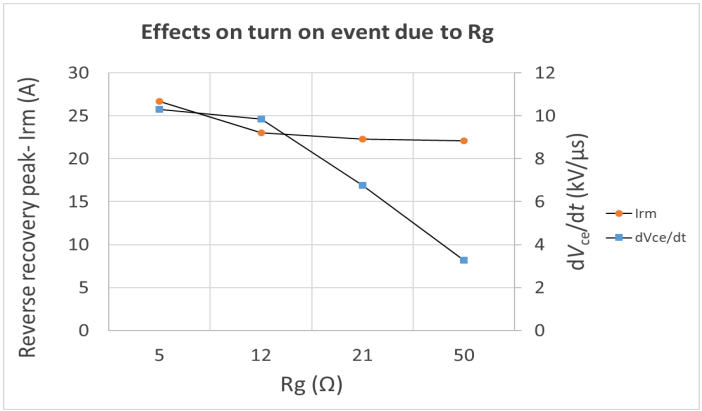


Figure 13 Effects on turn on event due to gate resistance (Full-Si IGBT) at V_{dc} = 400 V, I_c = 20 A and T_j = 25°C

6.2. Turn Off

6.2.1 Effect of Device Current

With increasing current, the losses due to all regions increase shown in Figure 14. The increment in losses is simply because of the higher value of current. The first region of turn off event, capacitive region, is determined by the rate of change of voltage $\frac{dV_{ce}}{dt}$. If $\frac{dV_{ce}}{dt}$ is higher, it means that the junction capacitance of the freewheeling diode discharges at a faster rate and that the diode is able to take the load current faster.

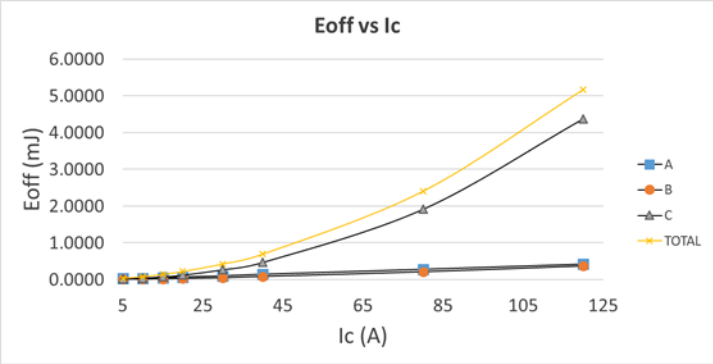


Figure 14 Turn off energy arising from its different root causes as a function of device current (Full-Si IGBT) at Vdc = 400 V, Rg = 14.6 Ω and Tj = 25°C A = capacitive, B = commutation, C = tail current

Here, with increasing current, there is no significant change in the rate of change of voltage as shown in Figure 15; therefore, the losses due to region A increases by small margin. This small increment is due to the beginning of integration from higher value of current.

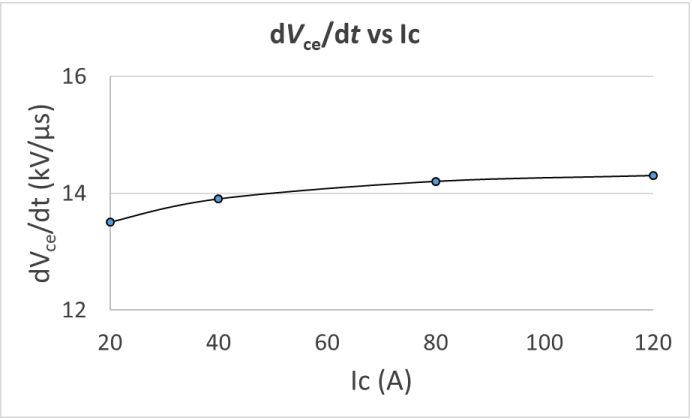


Figure 15 Variation in rate of change of voltage for different current levels (Full-Si IGBT)

It is clear from the Figure 14 that the turn off losses in an IGBT has a big part of losses coming from the tail current.

6.2.2 Effect of Junction Temperature

As temperature increases, the total turn off losses increases as shown in Figure 16. This happens because of the decrement in threshold voltage with increase in temperature so the time taken for the gate voltage to go lower than threshold voltage increases. This leads to slower switching speed; therefore, the power losses due to region B increases with increasing temperature.

The losses due to part C increase with increase in temperature as minority charge carriers take longer to recombine at higher temperatures.

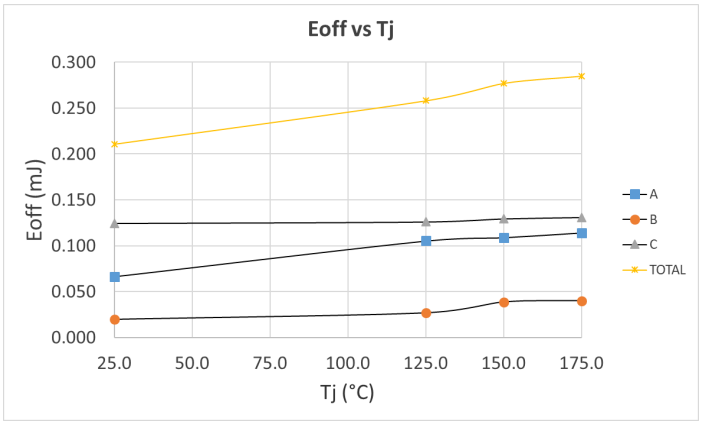


Figure 16 Turn off energy arising from its different root causes as a function junction temperature (Full-Si IGBT) at V_{dc} = 400 V, R_g = 14.6 Ω and I_c = 20 A A = capacitive, B = commutation, C = tail current

6.2.3 Effect of DC link Voltage

For higher DC link voltage, the turn off losses due to region A, B and C increase. This is because for higher V_{dc}, integration starts from higher value leading to increment in losses.

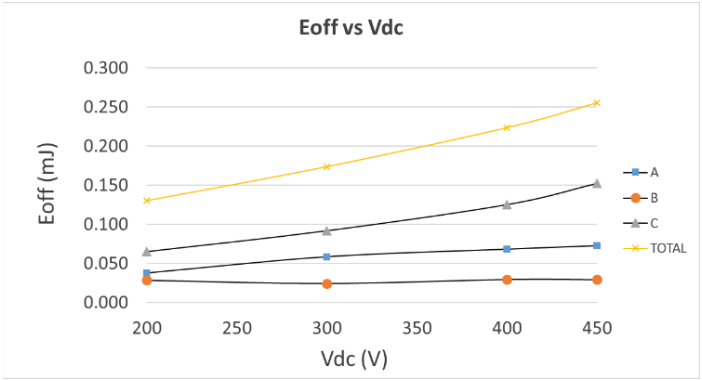


Figure 17 Turn off energy arising from its different root causes as a function DC link voltage (Full-Si IGBT) at V_{dc} = 400 V, R_g = 14.6 Ω and I_c = 20 A A = capacitive, B = commutation, C = tail current

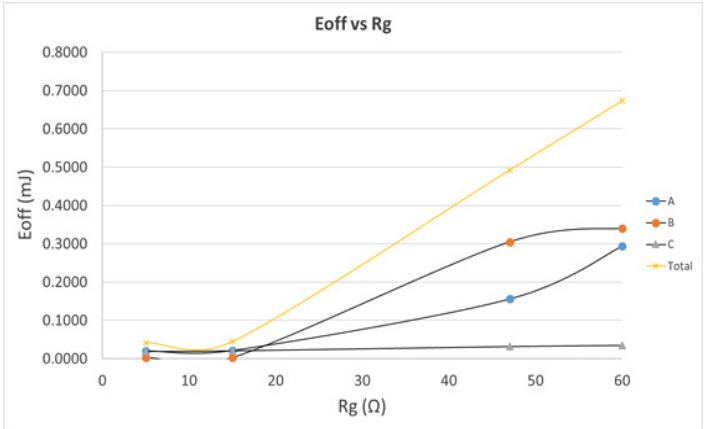


Figure 18 Turn off energy arising from its different root causes as a function gate resistance (Full-Si IGBT) at V_{dc} = 400 V, T_j = 25°C and I_c = 20 A A = capacitive, B = commutation, C = tail current

6.2.4 Effect of Gate Resistance

If higher value of gate resistance is chosen, then the rate at which device switches reduces. For higher gate resistance, $\frac{dV_{ce}}{dt}$ reduces and charging of junction capacitance is slower. This leads to increase in losses due to region A. Because of slower switching speed, the rate at which current falls down decreases leading to increment in losses due to region B. Since the tail current is specific to the IGBT; therefore, the losses due to part C do not change with the change in gate resistance [12].

7. Comparison of Switching Energies for the Considered Devices

Following the discussion of section 6, it is understood that all parameters affect both the devices in certain way. Figure 19 shows the turn on energies as a function of current at the same commutation speed with gate resistance of 5 Ω. At 30A, turn on energy reduces by 38% for Hybrid-Si/SiC IGBT. This is mainly due to the usage of unipolar Schottky SiC diode which doesn't have reverse recovery, leading to reduction in E_{on} by such a big margin.

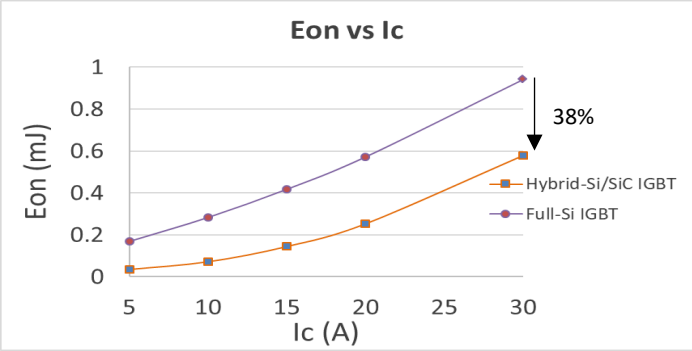


Figure 19 Turn on energy as a function of device current for Full-Si IGBT and Hybrid-Si/SiC IGBT at Rg = 5 Ω, Vdc = 400 V, Tj = 25°C

Figure 20 shows reverse recovery of the diode E_{rec} for different current levels. In the case of Full-Si IGBT, E_{rec} increases with increase in current. However, for Hybrid-Si/SiC IGBT, there is no significant change in E_{rec}, owing to the fact the losses for SiC diode do not originate from reverse recovery but due to the removal of charge carriers from junction capacitance.

8. Application Example

The commutation-speed based method can be applied to any application for the determination of the split-up of power losses at the system level. If the source contributing the most to the total losses is known, then steps can be taken in that direction to improve the efficiency. This method can help in the selection of most suitable device for the specific application. The essence of this method is demonstrated by applying it to previously discussed Boost PFC topology. The specifications of the implemented boost PFC are given in Table 2.

The power losses in this application are due to the rectifier stage, power switches used in boost stage and the losses due to passives. The losses in the power switches are modelled using the behavioral approach as discussed in section 2. Firstly, boost PFC with Full-Si IGBT is simulated in MATLAB where switching based simulation is performed i.e. losses at every switching instant is calculated, summed up and then averaged. This indicates that the simulation follows the same approach for loss calculation as discussed in section 3.

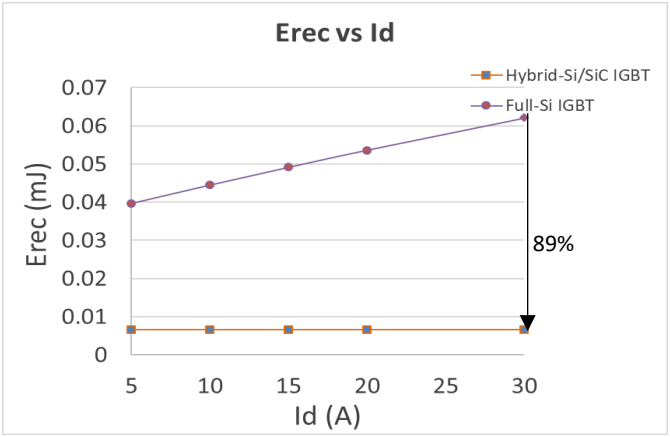


Figure 20 Reverse recovery energy as a function of device current for Full-Si IGBT and Hybrid-Si/SiC IGBT at Rg = 5 Ω, Vdc = 400 V, Tj = 25°C

Table 2 Parameters of the simulated boost PFC topology

Parameters	Values
V_{ac}	230V
P_o	3.3kW
V_{dc}	400V
L	263.8uH
C_o	1300uF
f_{sw}	80kHz
R_g	12Ω

Figure 21 shows the contribution of different sources of power losses in the boost PFC which are obtained from the simulation. It is seen in the Figure that there are significant losses due to the rectifier and passives in the boost PFC. However, to reduce losses of rectifier, the appropriate option is to use a different topology having less number of diodes in the rectifier stage, which will reduce the conduction losses in the rectifier.

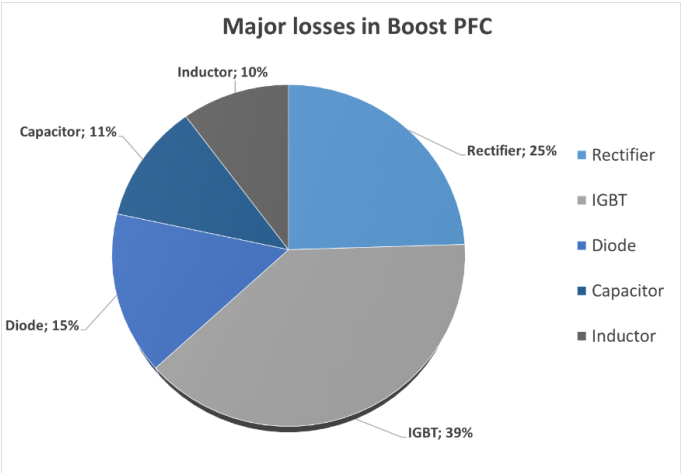


Figure 21 Average power losses in a boost PFC with Full-Si IGBT obtained from simulation at R_g = 12 Ω, V_{dc} = 400 V, P_o = 3.3kW and f_{sw} = 80kHz

The major contribution of losses is coming from the active switch i.e. IGBT. With just this knowledge, it is not easy to

determine the efforts that should be taken for the reduction of the power losses. At this point, the commutation-speed based method can be helpful as by splitting the switching losses, one can get a deeper insight into the losses that can help in the identification of the major cause behind the losses. This information can be used to figure out the steps that may be taken to reduce the losses.

This methodology is implemented on the considered two devices, the losses due to rectifier and passives are not considered further and the entire focus is on static and dynamic losses of the semiconductor switches.

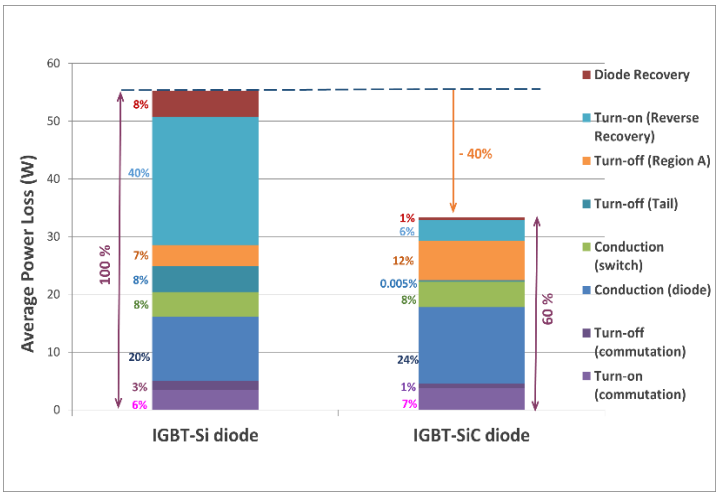


Figure 22 Split-up of the average power losses of the boost PFC for Full-Si IGBT and Hybrid-Si/SiC IGBT at R_g = 12 Ω, V_{dc} = 400 V

Figure 22 shows the split-up of average power losses for the considered devices according to the regions described in section 4. The contribution of each of these regions help in the assessment of the impact that each region has. Full-Si IGBT is taken as the reference and is marked as 100%. The contribution of all the regions are shown in the Figure as percentages of the total loss of Full-Si IGBT. In Full-Si IGBT, about 40% of power losses are coming from the reverse recovery part of the freewheeling diode (FWD) during turn on event. As discussed earlier, the reverse recovery losses depend on the type of FWD and replacing Full-Si IGBT with Hybrid-Si/SiC IGBT, the losses due to turn on region B decreases by 84% as there is no reverse recovery in a SiC diode and only capacitive losses exist. Besides, a reduction of 89% is observed in the losses due to the diode recovery because there are no excess carriers in a SiC diode. The other major part of losses is due to the tail current, but tail current cannot be eliminated in an IGBT. The only way to eliminate the losses due to tail current is to replace the IGBT with a MOSFET. Another important conclusion that can be drawn from Figure 22 is that conduction losses are slightly greater for SiC diode compared to Si diode. This is because of the forward characteristics of SiC diode.

On the other hand, it is observed that the turn off power losses due to region A are higher for Hybrid-Si/SiC IGBT as compared to Full-Si IGBT. This is mainly because of the junction capacitance of the diode which is bigger for SiC diode so it's discharging takes longer, leading to higher power losses. This is shown in Figure 23 where under same conditions, Hybrid-Si/SiC IGBT takes 43.2 ns for the discharge of junction capacitance of diode where as Full-Si IGBT takes 36.6 ns.

Figure 24 shows the performance of both the devices for boost PFC at different power levels and it is seen that the efficiency of Hybrid-Si/SiC IGBT is greater than the Full-Si IGBT which is mainly due to the reduction in the losses in diode recovery and reverse recovery part during turn on event. For output power of 3.3kW, efficiency increases by 0.7% for Hybrid-Si/SiC IGBT.

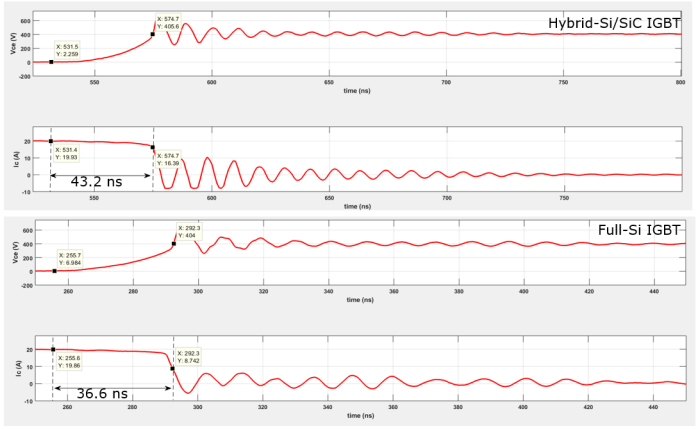


Figure 23 Turn off current for Full-Si IGBT and Hybrid-Si/SiC IGBT

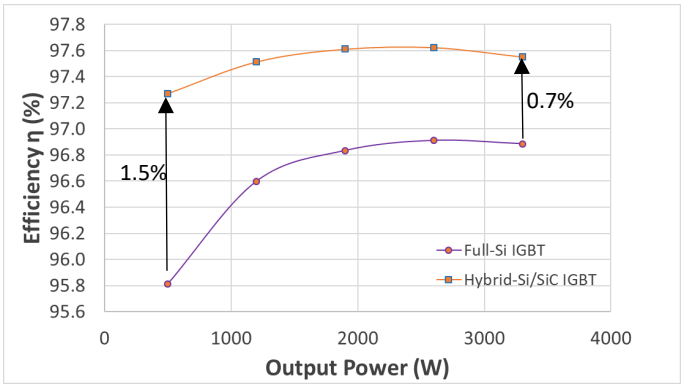


Figure 24 Efficiency comparison of considered devices for different power levels at V_{dc} = 400 V, R_g = 12 Ω, f_{sw} = 80kHz

9. Conclusion

Traditional power loss calculation gives limited insight into the root cause of the power losses. The commutation-speed based method was applied for the calculation of power losses in the boost PFC of an on-board charger. In this paper, the total switching losses for two discrete devices were split up into their root causes e.g. tail current, reverse recovery. The two combination of devices were IGBT with Si diode and IGBT with SiC diode. The split-up of the considered devices was studied in detail as how the variation in device current $I_{c/f}$, DC link voltage V_{dc}, junction temperature T_j and gate resistance R_g affect the losses. Static and dynamic characterization measurements were performed to model the two devices behaviorally which were then used in boost PFC to calculate the power losses at the system level. It was found that the losses in the switch reduce by 40% just by replacing Si diode with SiC diode for the same IGBT. This reduction in the losses is due to the decrement in the losses in diode recovery and reverse recovery during turn on event indicating that the replacement of just the diode from Si to SiC reduces the power losses by considerable margin. Hence, it can be concluded that this

methodology can be applied to any application for the identification of ways to reduce power losses and eventually leading to improvement in efficiency.

Conflict of Interest

The authors declare that there is no conflict of interest

Acknowledgement

We would like to express our gratitude to Rafael Garcia for sharing his knowledge during the course of this research work

References

[1] Gebhardt, Fabian, Hauke Vach, and Friedrich W.Fuchs, "Analytical Derivation of Power Semiconductor Losses in MOSFET Multilevel Inverters", Power Electronics and Motion Control Conference (EPE/PEMC), 2012 15th International. IEEE, 2012.

[2] Ajay Poonjal Pai Poonjal, Tomas Reiter and Martin Maerz, "A New Behavioral Model for Accurate Loss Calculations in Power Semiconductors", PCIM Europe 2016; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management; Proceedings of. VDE, 2016.

[3] Ajay Poonjal Pai, "Impact of Silicon Carbide based Power Modules on Mission Profile Efficiency of Automotive Traction Inverters", PhD Thesis to be published at Friedrich Alexander Universität Erlangen, (Internally available at Infineon).

[4] Ajay Poonjal Pai, Tomas Reiter and Martin Maerz, "Efficiency Investigation of Full-SiC versus Si-based Automotive Inverter Power Modules at Equal Commutation Speed", PCIM Europe 2018; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management. VDE, 2018.

[5] Infineon, "HybridPACK Drive, F2660R08A6P2 Power Module with 750V 660A EDT2 Chipset, Hints for Efficient IGBT Switching", Application Note, 2015, (Internally available at Infineon).

[6] Kihyun Lee, Kyungsub Jung, Seunghoo Song, Yongsug Suh, Changwoo Kim, Hyoyol Yoo and Sunsoon Park, "Analysis and comparison of high power semiconductor device losses in 5MW PMSG MV wind turbines", Power Electronics Conference (IPEC-Hiroshima 2014-ECCE-ASIA), 2014 International. IEEE, 2014.

[7] Ajay Poonjal Pai, Tomas Reiter and Martin Maerz, "An Improved Behavioral Model for Loss Calculations in Automotive Inverter", EEHE 2016 Wiesloch; Proceedings of 2016.

[8] Dennis Ward, Iqbal Hussain, Carlos Castro, Andreas Volke, Michael Hornkamp, "Fundamentals of Semiconductors for Hybrid-Electric Powertrain", Published by Infineon Technologies AG.

[9] Stefan Linder, "Power Semiconductors", EPFL Press, 2006.

[10] Infineon, "TRENCHSTOP™ 5 H5 (High Speed 5) IGBTs", Datasheet

[11] Markus Hermwille, "Gate Resistors-Principles and Application", Semikron International Application Notes.

[12] B. Maurice, L. Wuidart, "Drive Circuits for Power MOSFETs and IGBTs", ST Microelectronics Application Notes.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 268-281 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Strategies of the Level-By-Level Approach to the Minimal Route

Nikolay Starostin[*,1], Konstantin Mironov[1,2]

[1]*Institute of New Materials and Technologies, Ural Federal University, 620002, Russia*

[2]*Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, 450008, Russia*

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 07 February, 2019
Online : 20 February, 2019

Keywords:
Optimization
Minimum route
Traveling salesman problem
TSP
CNC

ABSTRACT

The task of optimal path planning for drilling tool with numerical control is considered. Such tools are used in the production of printed circuit boards. The algorithm modification of level-by-level route construction for the approximate solution of the traveling salesman problem is discussed. Bypass objects can be specified either as a table of distances or values represented by a symmetric matrix, or as Cartesian coordinates (in applied cases of using numerically controlled equipment). The algorithm was tested on many different examples. As a result of the calculations performed for examples from the TSPLIB library (defined through a full distance matrix or Cartesian coordinates) with a dimension of up to 100 cities, the ability of the algorithm to construct the optimal route was confirmed. For examples from other sources or for artificially constructed ones (up to 130 objects), in the testing of the algorithm, the declared record values of the route minimum length were also achieved or improved.

1. Introduction

This paper is an extension of the article presented at the IEEE International Symposium on Signal Processing and Information Technology [1]. Initially, the problem of the rapid acquisition of an optimal route for moving the tool between processing zones on computer numerical control (CNC) equipment was considered, which is close in form to the symmetric traveling salesman problem with the Euclidean metric. For a limited number of processing zones, a solution [2, 3] belonging to the category of composite methods according to classification [4] was proposed. Construction of a convex hull, use of branch and boundary method elements, level-by-level analysis and route options design. This made it relatively easy to synthesize the optimal result for a small number of processing objects. Subsequently, the need to increase the effectiveness of this approach was revealed, where an important role, just like in the branch and boundary method, is played by a successful choice of upper bounds for the solution for which heuristic algorithms are used [5].

In the process of developing the required approximate algorithm, it was natural to use state of the art techniques from the exact method [3], while somewhat broadening the general formulation of the problem. Namely (in terms of the traveling salesman problem), to find the shortest closed route passing through n cities, visiting each city exactly once. The distance between pairs of cities is determined by a symmetric, not metric in the general case, matrix of real numbers C=||ci,j||. In [6] a natural version of the transition to the approximate method of searching for the minimum route in this formulation is presented, while the basic procedure, the basic calculation schemes and possible directions of development are described. In [7 – 9], the calculation procedures are detailed and studied, the results of testing are given, along with options for constructing the algorithm based on different schemes from the combined levels. For all benchmark tests (examples from [10] up to 100 cities inclusive, given by a complete table of distances or by Cartesian coordinates), optimal routes are obtained. Also, for all of several dozen random tests, with reported results (dimension in the range of 30-60 cities), for example, as in [11], record values of the minimum length of the route were reached, or these figures improved.

2. Basic procedure and trivial calculation scheme

If we rely on the algorithms' classification from [12], then the level-by-level approximation (considered and tested in [7 ,9], respectively, for the matrix and coordinate representation) can be attributed to the symbiosis of "tour construction algorithms" and "monotonic algorithms for improving the tour." Specifically, such one when the procedure for identifying improvements to the tour is involved at different levels of the multi-stage building process. According construction details, the proposed algorithm is closest

[*]Corresponding Author: Konstantin Mironov, mironovconst@gmail.com

to a truncated combination of three algorithms from the classification [12]. This is **C (arbitrary connection)**. *«Let (i_1, i_2, …, i_n,) be an order of vertices. Form a two-member tour (i_1, i_2). The vertex i_l (l = 3, 4, ... n) is to be included in the (l—1) -segment (pod) in such a place that the increase in the tour ... is minimal».* **CR (most remote connection)**. *«Act like in C. At step l (l = 3, 4, ... n) in the (l—1) -st tour include the vertex i_l, for which the minimum distance to already included vertices is the largest among all vertices not yet included».* **CE (most economic connection)**. «*Act like in C. At step l (l=3, 4, ... n) in the (l-1) -st tour include the vertex i_l, for which the increase in the tour is minimal among all the vertices not yet included*. However, this combination emphasizes somewhat different points. It should also be noted that the proposed algorithm is multi-pass. Conscious departure from the seemingly natural use of coordinates for CNC equipment should also be underscored. This is done because different metrics can be used for different types of equipment and, accordingly, of technological processes (calculations of distances between treatment zones), therefore for the sake of greater generality we will not rely on the advantages of using metric space.

The essence of the basic algorithm is as follows. Let i1, i2, ..., in be a certain order of processing zones, vertices or cities in terms of the traveling salesman problem, where for definiteness i1 and i2 are the most distant cities from each other. Build a tour from i1 to i2 and back, i.e., a two-member closed tour (i1, i2, i1). Next, include the city ik (k = 3) in the existing two-member tour. This can be done in 2 ways: either (i1, ik, i2, i1) or (i1, i2, ik, i1). Since they are equivalent, we will explain the process using the first one as an example. So, for each of the two three-member tours, it is necessary to perform consecutive actions starting from the step l = 4. In step l (l = 4, 5, ..., n), in each of the available (l — 1) -member tours, include one of the remaining cities il (l = 4, 5, 6, ..., n) in that place of the tour, where it will provide a minimum tour increase based of the inclusion of this city. At the same time, the city of inclusion must be such that the value of the minimum increment of the tour from its inclusion is the maximum among the similar values of the minimum increments from other cities of the potential inclusion. Thus, at the last step (l = n) we obtain two complete routes. To fix one of them with the smallest length (in general, their length is the same, and the routes themselves coincide, the difference is only in the direction of the detour in the forward or reverse direction). Thus, the first of the approximate routes (Hamiltonian cycle) is obtained, which corresponds to its progenitor - a three-member tour i1, ik, i2, i1. (k = 3). Then go back to the original two-member tour (i1, i2, i1) and perform similar calculations for the remaining cities ik (k = 4, 5, ... n). As the result, we will obtain n-2 approximate route options. Each route is determined by uniform constructions from its initial three-member tour i1, ik, i2, i1, (k = 3, 4, ... n). We call the routes formed in this way from the original three-member tours, routes of level 0 and, for convenience, assign them indices (00, 01, 02, ...) in accordance with the order of their length increase.

It can be noted that the level 00 route (s), with any significant number of cities, rarely reach optimum or are close enough to it. Therefore, it is required to continue similar constructions in order to obtain level 1 routes. They are similarly constructed for each of the n-2 initial three-member tours i1, ik, i2, i1, (k = 3, 4, ... n). At the same time, at least one of the new routes (for each of k = 3, 4, ... n) will be no longer by construction than its related route of level 0.

Let us explain this calculation by the example of constructing the first of the routes for level 1. In the initial tour (i1, ik, i2, i1), (k = 3), include the city im (for definiteness m = 4) alternately between pairs of adjacent cities. Accordingly, obtain three 4-membered closed tours (i1, im, ik, i2, i1), (i1, ik, im, i2, i1), (i1, ik, i2, im, i1). In step l (l = 5, 6, ... n), in each of the available (l — 1) -member tours, include one of the remaining cities il (l = 5, 6, ..., n) in the place of the tour where a minimum tour increase will be provided from the inclusion of this city. At the same time, the city of inclusion must be such that the value of the minimum increment of the tour from its inclusion is the maximum among the similar values of the minimum increments from other cities of the potential inclusion. Thus, at the last step (l = n) we get three complete routes. Fix one of them with the shortest length. This is the first of the approximate level 1 routes. Together with it, the initial 4-member tour is fixed as well (the value of m, k and the location of the corresponding cities in the tour).

To form the remaining routes of this level (with k = 3), we need to return to the original three-member tour (i1, ik, i2, i1) and perform similar actions with other valid values im (m = 5, 6, ..., n), as in the construction of the first route of level 1. As a result, for the value of k under consideration, we obtain n-3 Hamiltonian cycles, i.e., approximate routes of level 1.

To form the remaining routes of this level (n-3 for each new k, where k = 4, 5, ..., n), we need to return to the original three-member tours (i1, ik, i2, i1). Perform similar actions for each of k with admissible values of im, as in the construction of the first group of routes of level 1. As a result, for each of k (k = 4, 5, ..., n) we get n-3 full routes of level 1. All received routes in each group (fixed value k) are assigned indices (10, 11, 12, ...) in accordance with the order of increasing their lengths (several routes can have the same index if their lengths are equal to each other).

Similarly, routes of levels 2, 3 and subsequent are built, respectively, from 4, 5 etc. member source tours identified at the previous level.

But if nothing restricts this process, then we will have an avalanche-like growth of calculations. At the same time, when moving inward through the levels, a situation similar to the “thermal modeling” algorithms [12, 13], may be observed, where the process of improving the complete route is attenuated. For greater certainty, in [9], based on the accumulated statistical data, examples of different dimensionality of the level number, beyond which constructions are hardly advisable, are roughly indicated. When the level-by-level constructions are terminated according to the conditioned criterion, then in accordance with the multipass algorithm, we must again go to the two-member closed tour (i1, i2) for successive inclusion of the vertex ik (k = 4, 5, ... n), as in the sample presented above (for the case k = 3). As a result, the route, the smallest of all constructed, is determined. It is also a potential contender for correspondence to the optimal route. This correspondence is consistently confirmed when testing the examples of a relatively small (up to 60 elements) dimension. Additionally, in practice, for achieving records from the original tripartite tours (i1, ik, i2), in tests from [10, 11, 14] with a dimension of 25-60 vertices, it was sufficient to perform calculations not for all but only for several vertices ik where the values of index k correspond to those inclusion vertices that form zero-level routes with lower indices [7, 8]. In addition, for an overall reduction in the calculation of tests with a dimension of 60-100 vertices, extended schemes can be used, so that to reach a

record we would also be limited to calculations with only lower indices from the starting extended level [9].

In general, starting from this basic procedure with a trivial calculation scheme, it is necessary to form strategies rational for practical use to create complete routes and check the criteria for stopping calculations.

3. Possible directions in organizing calculations.

First, we return to the problems of CNC machine tools. Drilling holes in printed circuit boards is one of two sources of TSP instances that are associated with practical application [16]. From an applied point of view, local search algorithms are best for solving the traveling salesman problem [17]. Nevertheless, practical problems of this kind can bring certain difficulties for local search, even with a small number of vertices. So, for example, one of such effective algorithms with the Lin-Kernigan neighborhood implemented in the Concord program [17], does an excellent job with large-dimension tests from [10], but gives, for example, a fragment of a printed circuit board with a dimension of 15 vertices, which is not an important result (approximately 103% from optimum [1]). To obtain the optimal route (Figure 1 on the right), the program needs to change the default settings, which may not help at all in more complex cases. In addition, metaheuristic methods are very popular for practical problems. [18]. They do not guarantee that the best solution to the problem will be achieved, but with a successful setup they can find a "good enough" result. Their disadvantage is the presence of a different number of control parameters [20], which makes it difficult to set up algorithms for solving specific problems.

Now let us consider possible approaches for the organization of calculations in the framework of our basic algorithm for the level-based approximation.

3.1. Calculation on the basis of not deteriorating routes

Theoretically, each of the n-2 initial 3-member tours i1, ik, i2, i1, (k = 3, 4, ..., n), from which level 1 routes are formed, can be the progenitor of the optimal route. That, however, can no longer be said about each of the 4-member (and subsequent) tours, from which routes of level 2 and further are formed. Therefore, after constructing routes of level 1 (for each of the n-2 three-member tours), the route (s) with the index zero (10) is (are) allocated to each of the received n-2 groups and the corresponding four-member tour is fixed for them, from which the subsequent construction will go (routes with a different level index and the corresponding components are not considered further). Level 2 routes are similarly constructed on this base (level 2), after which only route (s) with index zero (20) and so on is (are) also considered (that is, at subsequent levels everything related to non-zero indices routes is not participating). Thus, routes of non-zero level J (J = 2, 3, 4, ...) are constructed similarly to routes of level 1, from their initial (J + 2) -fold tours, which correspond to full routes of levels (J-1) 0 . The whole process of level building can be conditionally represented by the following scheme:

(1) $\boldsymbol{level\ 0_i - level\ 1_0 - level\ 2_0 \ldots (i=0,1,2,\ldots)}$

The shortest length among all level 1 routes will be the route with a zero index (level 10), which is the result of the initial embedding of a certain city iz ($z \neq 1,2, k$) into a certain place of the tour (i1, ik, i2, i1). By construction, its length cannot be greater than the length of the "parent" route of level 0, resulting from embedding in the tour (i1, i2, i1) of city ik. The same applies to all derived routes with a zero index of subsequent levels, i.e., the best routes of the level cannot be longer than their ancestors. Moreover, when building at the initial levels (when moving from a lower level number to a larger number), the length of the descendant route decreases; as practice shows, this is a frequent case. However, as the level numbers increase, this happens less frequently and then fades out.

Therefore, the process of constructing routes in ascending level numbers continues, approximately as in "thermal modeling" algorithms, until a certain equilibrium is reached, understood as "no significant deviations from the found tour length are observed for a long time" [12, 13], i.e., until the termination of reducing the current length of the route. After that, the level constructions in this direction are interrupted. In the end (when the calculation is stopped in all available directions) the route is determined, the smallest of all constructed. It is also a potential contender for compliance with the optimal route. In [6, 7], the main results of testing in the framework of the considered scheme (1) for several dozens of tasks with cities given by a symmetric table of distances up to 60 objects in dimension are presented. At the same time, in all calculations, the reported record results were obtained or improved.

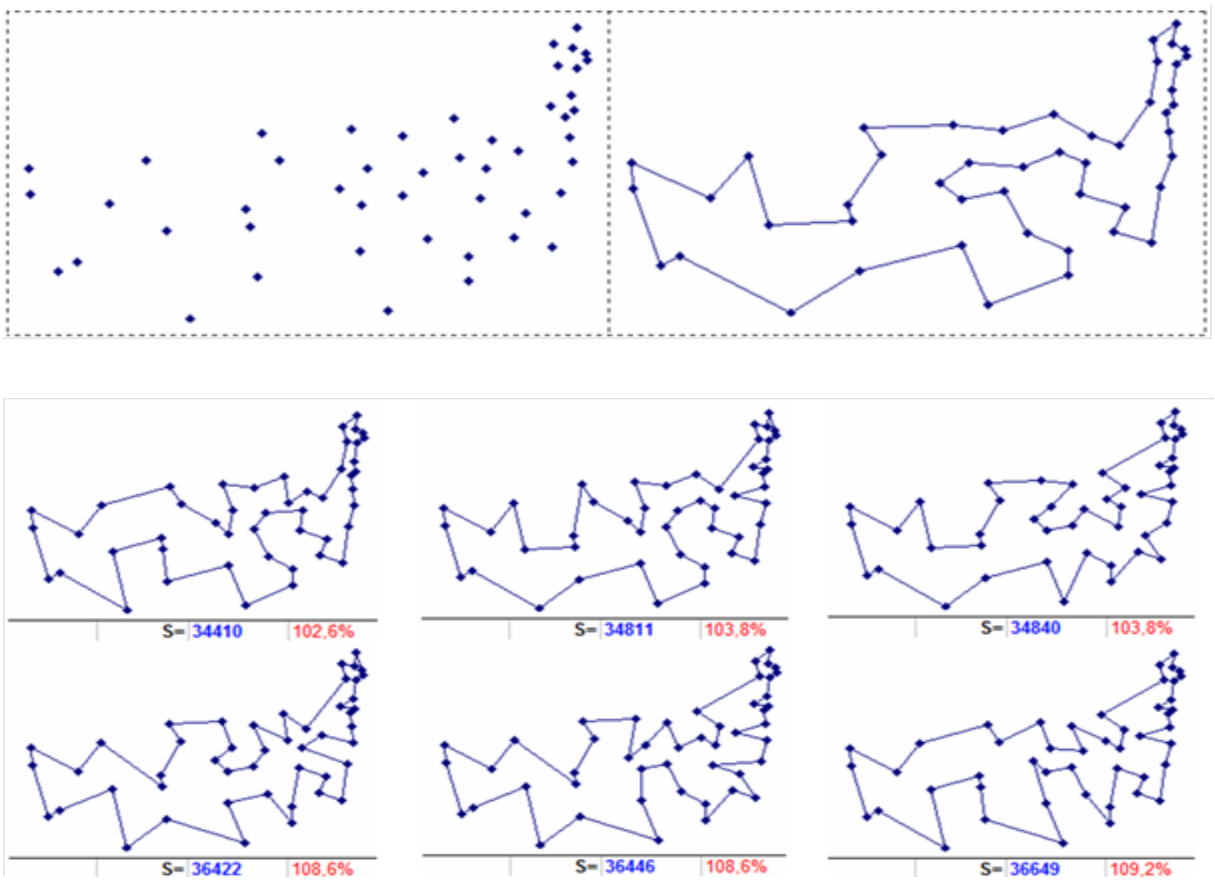
To explain the essence of the material presented in [1], a simple example was selected with 29 bypass objects from the TSPLIB library [10], and to demonstrate the results of innovations, a test (file xqf131.tsp) measuring 131 elements from the VLSI TSPs collection [20] created in based on SBIS data sets investigated at the University of Bonn.

Here, as a simple illustrative material, we take a classic sample containing 48 objects (att48_d.txt file) from a library with TSP data [21]. This is one of the key examples put on the cover of a famous book [22], which was used in it to illustrate existing approaches to solving the problem in question. To demonstrate the results of innovations, we will again use the same test of 131 elements.

So, let us explain the organization of the process for finding the minimum route on the selected example with 48 US cities (att48_d.txt file), given by the complete distance table. For graphical representations of the constructed routes, the Cartesian coordinates of cities from the att48.tsp [10] file will be used (in Fig.1, coordinates are on the left, and the known optimal route with a length of 33551 is on the right).

If we number the rows of the original distance table starting from 0, then the most distant cities will have numbers 3 and 16. Route level 00 (Fig. 2 upper left) built from the initial two-member tour (by cities 3, 16, 3) has a length of 34410 and a percentage ratio of 102.6% with the optimal route (hereinafter, percentages are indicated to the nearest tenth). The same figure shows the routes of level 01, level 02 (continuation of the first row of Fig. 2) and in the second row - the three longest routes of level 0 (which have the largest indices).

Sample construction 1. Route level 00 (Fig. 2 left upper) in one copy. It assumes the inclusion of the city with the number 30 in the two-member tour (3, 16, 3) (other cities with similar embedding give a greater length of the resulting route of the zero level, and, accordingly, such routes will have a larger index.



The route of level 10 (00–10), built from the initial three-member tour (3, 30, 16, 3) is also one, its length is 33961 (Fig. 3 on the left). It implies embedding city 46 in the tour (3, 30, 16, 3) between cities 16 and 3.

Level 20 routes (00–10–20) are already 15, their length has not improved, and has remained the same (33961) as at the previous (parent) level 10. Each of them involves embedding one of the 15 cities: 5, 6, 7, 9, 14, 17, 18, 26, 27, 29, 34, 36, 37, 43, 44 into the initial tour in a certain place (3, 30 , 16, 46, 3). Next, for each newly formed tour routes of level 30 (00 - 10 - 20 - 30) are built, of which there are already about 152, which also do not show the change in the length of the best tour. And only at level 40 there is a decrease in length in certain directions (Fig. 3 in the center and to the right).

First, we shall consider the direction corresponding to the shorter route length (33614 Fig. 3 on the right). Let us call it “branching 1”. When building routes of level 50 from it, there will be several of them again, i.e., 22, from each of which at level 60 will be somewhere along 23-27 similar new ones, and their length does not change (it remains equal to 33614). The situation is similar at the level of 70 and 80. That is, there is a circumstance that “for a long time no changes are observed” of the lengths of routes with zero indices and therefore it is necessary to stop further constructions in this direction.

For the future, we should agree on the criteria for stopping calculations in the current direction. Let it be, for example, the following: “threefold repetition of the route length value with a zero index during the transition from level to level in the process of forming hereditary directions».

That is, in our case, calculations for this direction (branching 1) should be stopped after the formation of level 70 routes, despite the fact that we would reach the optimum (length of the route is 33551 cm. Fig. 1) already while calculating from level 90, continuing calculations for this direction. For example, when calculating from the initial closed tour 3, 1, 33, 7, 30, 16, 32, 19, 46, 31, 4, 3 (when you include city 40 between cities 33 and 7). Or from another source tour 3, 1, 33, 40, 7, 30, 16, 32, 46, 31, 4, 3 (with the inclusion of the city 19 between cities 32 and 46).

Let us now consider the second direction (“branching 2”), corresponding to the greater length of the route (33741 Fig. 3 in the center). When building from it, 4 routes of level 50 are obtained with a length that has changed to a value of 33614. From each of the corresponding four routes, the eight-member initial tours are obtained at level of 60, with about 21-24 new routes in them, while their length does not change (it remains equal to 33614). At level 70, another average of 23 routes is added. This is about 4 * 232 new calculation directions in grand total. And if during the formation of the level 80 routes there is no change in the length of at least one of them, then, according to the entered criterion, the calculations in this direction (“branching 2”) should be stopped, and therefore, in general, the entire “parent” direction of level 00 should cease. However, in reality this does not happen. When forming routes of level 80 from the initial closed tour 3, 40, 2, 8, 7, 30, 6, 16, 46, 3 (when you include city 32 between cities 16 and 46 - constructing at level 70), a route of length 33551 (Fig 1) results, which is an optimal one.

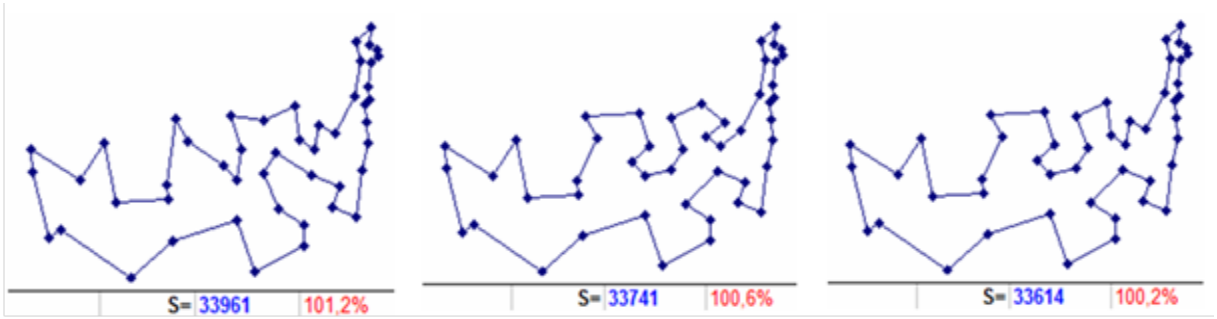
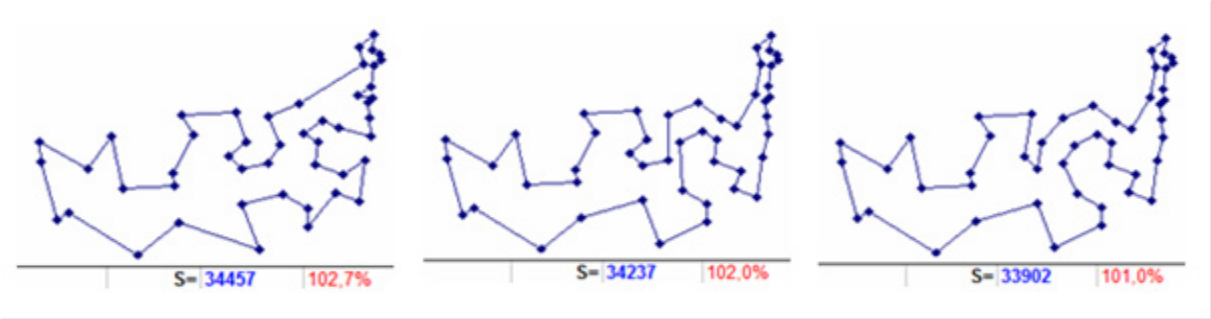


Figure 3. Route of level 10 (left), and of level 40 (center and right). Additionally shown are the length and deviation relative to the optimum.



Despite the fact that the result within the declared scheme (1) has been achieved, we will continue our constructions from the 0i ($i \neq 0$) levels in order to identify less labor-intensive directions. And such options are, for example, the following ones.

. This route is one copy. It assumes the inclusion in the two-member tour (3, 16, 3) of the city with the number 32. There are three routes of level 1_0 ($0_{24} - 1_0$), constructed from the original three-member tour (3, 32, 16, 3), their length is 34457 (Fig. 4 left). They suggest embedding in the tour (3, 32, 16, 3) cities 10, 19 or 34 between cities 16 and 3.

We will begin further constructions from city 10 (having a smaller sequence number). When calculating from the tour (3, 32, 16, 10, 3), the only route of level 20 is 34237 long (Fig. 4 in the center). It assumes embedding in the tour (3, 32, 16, 10, 3) city 43 between cities 16 and 10. By performing this operation, we also get a single route of level 30 with a length of 33902 (Fig. 4 on the right). This, in turn, is assuming that the only embedding in the tour (3, 32, 16, 43, 10, 3) shall be city 22 between cities 3 and 10. This embedding, in turn, results in obtaining an optimal route of length 33551 (Fig. 1). As a result, the entire calculation is significantly less time consuming than the calculations from level 00 in sample construction 1. The remaining two branches (those from cities 19 and 34) are more computationally expensive.

Here we shall note that it is possible to dramatically increase the number of ways to build optimal routes through the use of extended schemes for non-deteriorating routes.

In [8] it was shown that within the framework of the scheme (1), as a rule, there are quite a few options for constructing optimal routes; however, in some cases it is not possible to quickly reach the optimal result, therefore in [7], alternatively, extended schemes are proposed. They allow, for example, in problems with a dimension of up to 50 elements, to obtain an optimum often even at the initial state. Either they open new directions in which the optimal result is achieved in the absence of significant branching of the process, or at earlier than in scheme levels (1), which may already be more relevant for tasks in the range of dimensions, from 50 to 100 elements [9].

The first version of the extended calculation scheme:

(1) ***level (0 – 1)$_i$ – level 2$_0$ – level 3$_0$… (i=0,1,2,…)***

Here, level (0 - 1) represents some combination of level 0 and level 1 from the previous scheme. Full routes (Hamiltonian cycles) from level 0 are not built, and each of the n-2 source 3-member tours i1, ik, i2, i1, (k = 3, 4, ..., n) is converted into 3 four-member one, from which the full routes are already formed from the combined level (0-1). Here, they are presented in the number of (n 2) * (n 3) * 3, which is a broader basis for the formation of full routes of levels 20 and subsequent levels. At the same time, within the framework of the scheme (1) for the formation of level 20 routes, there is, as practice shows, in all directions (n 2) or several more closed source routes, that is, a small subset of the analog from the scheme (2).

Let us explain the scheme (2) in our current example. Among the many variants of the initial routes corresponding to the level (0 - 1) i, we take two (with some average level indicators of the full route length) and at the same time non-existent (unrealizable) within the framework of the scheme (1).

Sample of construction 3. Let the first of them be the original route that implies embedding into a two-member tour 3, 16, 3 pairs of cities 15 and 30 between cities 3 and 16. As a result, we have tour 3, 15, 30, 16, 3, and the only route from it is on level 20, involving the incorporation of city 37 between cities 30 and 15, with a total length 34410, corresponding to the upper left in Fig. 2

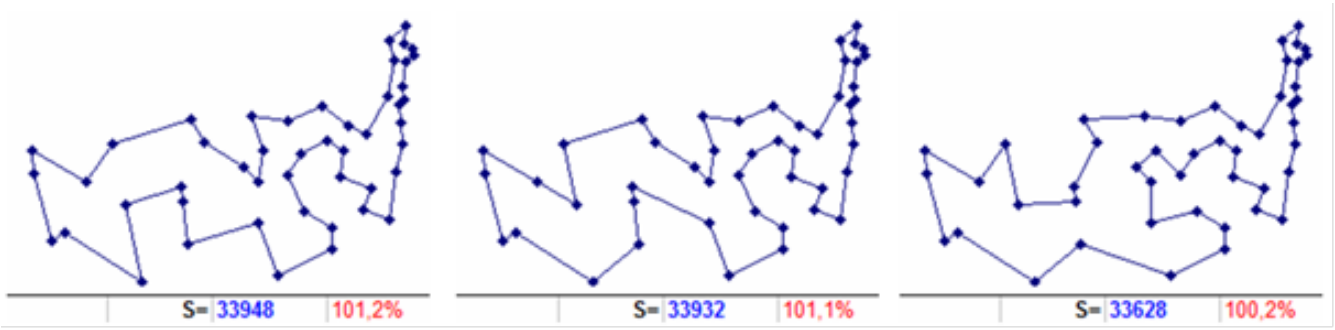


Figure 5. Intermediate routes of level 30 (left), level 40 (center), level 50 (right) as part of the implementation of the scheme (2). Additionally, the length and deviation relative to the optimum are shown.

In turn, it builds only one next-level route with a zero index (level 30), with an estimated integration of city 10 between cities 15 and 37 on tour 3, 15, 37, 30, 16, 3 with the result (length 33948) indicated in Fig. 5 left. Already from this position is also obtained the only route of level 40 with the expected inclusion of city 25 between cities 3 and 15 in tour 3, 15, 10, 37, 30, 16, 3, and with length of 33932 (Fig. 5 in the center). Further, the route of level 50 is also created in a single copy indicating the inclusion of city 12 between cities 15 and 10 in tour 3, 25, 15, 10, 37, 30, 16, 3, and with length of route 33628 (Fig. 5 on the right). And already from that, after the implementation of the specified inclusion, we immediately get the optimal route (Fig. 1 on the right). As a result, all transitions from the initial to the last level pass without a single branching in the process of construction.

Sample of construction 4. Let us take, as the second similar initial tour, a route with a supposed embedding in the two-member tour 3, 16, 3 pairs of cities 22 and 30 between cities 3 and 16. There is also no corresponding route within scheme (1), since embedding from level 1 with city 22 in the case of constructions from level 00 (or with city 30 in case of constructions from level 025) will have an index number of level 1 greater than 0, which is beyond the scope of the scheme. Within the framework of scheme (2), the corresponding full routes exist and coincide, their length is 35935 (Fig. 6, left). The calculation from tour 3, 22, 30, 16, 3 gives a single route of level 20, which implies embedding city 11 between cities 22 and 30 with a full route 34694, corresponding to the central route in Fig. 6. There are already two level 30 routes (Fig. 6 on the right), they suggest embedding cities 32 or 45 between cities 3 and 16 in tour 3, 22, 11, 30, 16, 3. Both routes with the specified embedding lead to an optimal result (Fig. 1 right).

Sample of construction 5. It is also possible to note the possibility of obtaining an optimum according to the scheme (2) among the directions existing within the framework of the scheme (1), but "rejected" by the criterion of stopping the calculations. For example, in the framework of scheme (1), according to this criterion, constructions are rejected from the direction (06–10–20–30) because of the fourfold construction at each level of the same complete route. Namely, the sequential inclusion in the two-member tour 3, 16, 3 the city 43 (3, 43, 16, 3), then the city 10 (3, 10, 43, 16, 3). After that, city 40 (3, 40, 10, 43, 16, 3) and finally, city 0 (3, 40, 10, 0, 43, 16, 3), each time they build the same full length route 33633 (100, 2% of the optimum of Fig. 7, in the center). In fact, there is a threefold repetition of the length of the route and, according to the criterion, the calculation stops.

Let us present another extended scheme similar to (2), but with three levels combined.

(1) ***level (0 – 1– 2)$_i$ – level 3$_0$ – level 4$_0$… (i=0,1,2,…)***

It provides even greater opportunities for obtaining record-breaking results due to the impressive "re-selection" component of the combined entry level. *(0 – 1 – 2)$_i$*.

Sample of construction 6. So, for example, thanks to such a scheme, from tour 3, 12, 10, 43, 16, 3 (which corresponds to the construction from the combined level (0 - 1 - 2)) in our current example of 48 cities, we immediately get the optimal route. This tour is not the only one on the level combined by the scheme (3).

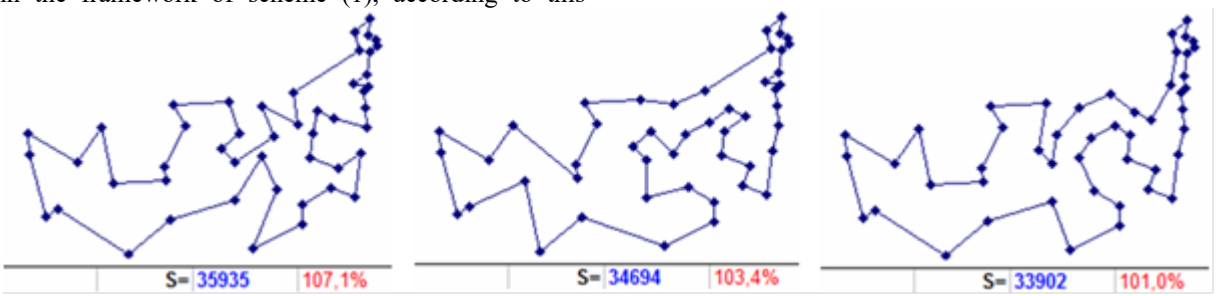


Figure 6. Intermediate routes of the "combined" level (0-1) i (left), level 20 (center), level 30 (right) as part of the implementation of the scheme (2). Additionally, the length and deviation relative to the optimum are shown.

3.2. Calculation based of the scheme with permissible deterioration of routes

Such a scheme is designed to provide flexibility in achieving record routes. This is a way to improve from a slightly worse calculation position (which is longer than the best route of the current level), which at the next or later levels can lead to a change in the current minimum in the considered direction of calculation. It is significant that this allows to avoid the situation "for a long time, no deviations from the found tour length are observed". This approach is implemented based on the use of index variation not only as previously at the zero level, but also at subsequent levels of construction. That is, in the general case, the calculation scheme can be as follows [16]:

(2) ***level 0$_{i0}$ – level 1$_{i1}$ – level 2$_{i2}$ … (i0, i1, i2, …=0,1,2, …)***

The total number of calculations here is much larger relative to the scheme (1). Therefore, from a practical point of view, one should artificially limit the potential excessive growth of the level-by-level calculations. And from this position, it is more expedient to introduce into (4) additional restrictions or to use some hybrid variants from schemes (1) - (3) and (4), for example let us consider the scheme (5), which is defined by the next three paragraphs.

If at the levels of J0, where J = 1,2,3 ..., only one route is observed (with index 0), then further constructions lead only from the initial (J + 2) -member tour for it, i.e., according to scheme (1).

If there are several routes with index 0, then further constructions will lead not only from the initial (J + 2) -member tours for them, but also from all other (J + 2) -member tours that are initial for the level routes with corresponding non-zero indices, i.e., as in scheme (4).

Moreover, in order to reduce computations, directions of calculations from non-zero indices, for constructions from which the full route with a zero index of the next level coincides with the parent route (that is, in the absence of a decrease in the length), shall not be considered,).

To illustrate the scheme with permissible degradations of routes, we turn to the already considered version of the calculation from construction sample 1, where for the first time instead of one, in the direction of 20 (in the direction 00 - 10 - 20), 15 routes appear with a zero index and not improved length of 33961 (Fig. 3 left). This is just the reason to use the scheme (5).

Sample of construction 7. So, we shall start constructing routes of the 3rd level. Embedding in a certain place of the initial tour (3, 30, 16, 46, 3) each of the identified cities corresponding to the level 2 zero index does not change the generated routes of the next level with the zero index (see construction pattern 1, level 30). According to the adopted scheme, we will construct from non-zero level indices. So embedding of city 15 between cities 3 and 30 (corresponding to the route with the last, i.e., the largest index of level 2 with an initial length of 35582) results in two routes of level 30 with a decrease in length to 33614.

Further, from the first of these routes, through the implementation of the supposed embedding of city 33 between cities 3 and 15 (on tour 3, 15, 30, 16, 46, 3), 25 routes of level 40 are obtained. In one of them, the embedding of city 32 between cities 16 and 46 (in tour 3, 33, 15, 30, 16, 46, 3) allows to obtain the optimal route which is several levels earlier than indicated in construction sample 1.

From the second route, an optimal tour is also obtained, but it is built one level further.

Sample of construction 8. A stronger result is obtained when constructing from level 06, where there is (before the start of the intended embedding of city 43 in the two-member tour 3, 16, 3) a full route (Fig. 7, left) 35143 in length. The route of level 10 from the implementation of such embedding is one, with length of 33633 (Fig. 7, in the center) and it involves embedding the city 10 between 3 and 43 (on tour 3, 43, 16, 3). When implementing it, we already get 10 level 20 routes and, according to scheme (5), along with them, it is possible to conduct calculations also from levels 2i (where $i \neq 0$). As a result of the calculation from level 27 (where the full route 34479 in Fig. 7 is on the right), when the city 22 is embedded between 3 and 10 (in tour 3, 10, 43, 16, 3), we immediately get the optimal route (Fig. 1 on the right).

It can be noted that in a similar way and with the same laboriousness (when building from level 2), optimum is obtained when calculating from the starting level 012 (the full potential route is no longer than 35808), with an estimated only integration of city 10 into the two-member tour 3, 16, 3.

4. Natural innovation in the basic calculation procedure

Having considered the main directions of calculations, and based on the statistics of the calculations carried out earlier [7–9, 16], we can draw some conclusions. For the considered possible directions of calculations in various test examples with dimensions up to 100 and a little more elements, it is always possible to obtain the optimal result with varying degrees of complexity. In this case, the optimum for each of the possible calculation schemes, as a rule, is achieved in many different ways. The least laborious route

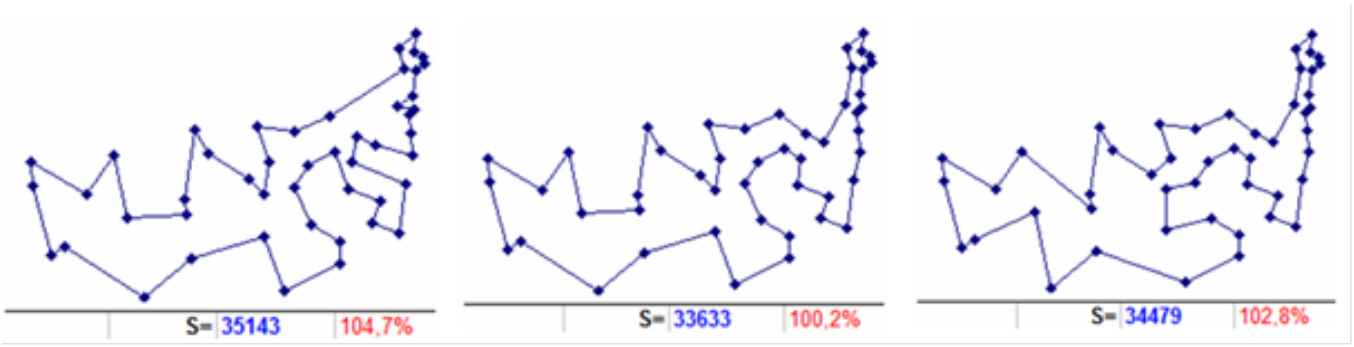


Figure 7. Routes corresponding to the constructions according to the scheme (5): level 06 (left), level 06-10 (center), level 06-10-27 (right). Additionally, the length and deviation relative to the optimum are shown.

(when comparing calculations from level 0 starting tours) was often the one where the optimal route is built at earlier levels. Less often - at later levels, but in the absence of a large number of branches. The results of experiments [7–9, 16] also show the following. The more extensive the base of the starting (initial) level, as is the case with the use of advanced schemes, the higher the probability of an "earlier" approach to optimum in individual calculations from starting tours. The total number of such exits is also growing.

There is a natural opportunity to significantly increase the number of zero level starting tours from the current n-2 options within the framework of the scheme (1). That will also increase the number of corresponding approximate level routes. The essence of such an extension is as follows. For convenience and certainty, the baseline calculation procedure used the directive choice of the first two cities (maximum distance from each other). From this pair, all subsequent constructions were carried out. If calculations would be carried out from each unique pair of source cities, then the number of starting two-member tours increases to n * (n-1) / 2, where n is the number of cities. Let us return to current example (48 cities) to clarify application of new starting conditions.

Samples of construction 9. In practice, probably, such a large number (in our new conditions 1128) of initial tours is not necessary. Let us see what happens if we restrict ourselves, for example, to considering only a few starting two-member tours, from which when forming routes of the zero level we get the smallest routes with zero index (i.e., the smallest among all possible routes of levels 00).

In our example, there are only 68 (out of 1128) two-member tours with a minimum route (at level 00) of a length of 33614 (as in Fig. 3 on the right). By constructing according to scheme (1), we will reveal among them those two-member tours, from which the optimal route will be constructed at the nearest of the subsequent levels, and accordingly the number of this level. Thus, directly from all possible three-member tours, being built from each of the 68 two-member tours, the optimal route is not immediately reached. But for the individual subsequent four-member tours the optimal route (Fig. 1 on the right) is already calculated. That corresponds to the constructions from level 10. The code of these constructions, corresponding to scheme (1), is contained in table 1; it indicates the sequence of obtaining the optimal route (length 33551).

Table 2 contains similar information, but this time from a small sample of the remaining starting two-member tours. In it, the best routes of zero levels (routes of level 00) exceed the minimum length of 33614 (that is, more than the tours from the previous table). The corresponding constructions for each of those two-member tours of scheme (1) also lead to the identification of the optimal route, with the implementation of inclusion from level 10.

The above tabular data suggests that there are no clear preferences when choosing a two-element tour from the available starting conditions in order to obtain a "short" calculation of the optimal route. For example, our initial starting two-member tour (3 16 3) from the most distant cities is with index 75 according to the criterion for the route length of level 00 almost exactly in the middle of the totality of 1,128 two-member tours (among the sequence numbers in ascending order from 532 to 574). From the point of view of building an optimal route (sample of building 1 - 8), it is not very fast relative to others. But the starting two-member tour 12 38 12 with the sequence number 1127 (Table 2 last line) with the length of the 00 level route equal to 35400 (Fig. 8 on the left) is almost perfect from the same point of view (like some others presented in both Tables).

And this is despite the fact that its starting indicators are next to the last ones in proximity to the optimum length of its best route among all 1128 results of levels 00. However, the optimal path is quickly and unequivocally (without a single branching) built from level 03 (basic length 35659 in Fig. 8 in the center). Specifically, first begins the inclusion in the two-member tour of the city 28 (12 28 38 12). As a result, a route with the length of 34603 is built (Fig. 8 on the right). Then, the only possible inclusion of the next level 10 is the insertion of city 32 (12 32 28 38 12) and, as a result, the already known optimal route is obtained (Fig. 1 on the right).

Table 1: Sequential Data

The sequence number of the original tour	The order of the cities for the starting two-member tour	The length of the level 0i tour(at the beginning of the direction for constructing the optimal route)	Order of cities of a three-member tour (*level 0_i*)	Length of a tour of level 1_0	Order of cities in a four-member tour, directly from which the optimal route is built (*level 1_0*)
3	1 30 1	35374 34110	1 **12** 30 1 1 **15** 30 1	33628 33628	1 **15** 12 30 1 —— // ——
8	1 43 1	35137 35131	1 **22** 43 1 1 **24** 43 1	33879 33879	1 **22 24** 43 1 —— // ——
22	6 37 6	34947 35613	6 **22** 37 6 6 **24** 37 6	33955 33955	6 **22 24** 37 6 —— // ——
39	17 37 17	35191	17 **24** 37 17	33955	17 **22** 24 37 17
45	22 27 22	34483	22 **24** 27 22	33955	22 24 **37** 27 22
53	27 37 27	34704 35769	27 **12** 37 27 27 **24** 37 27	34701 33955	27 **28** 12 37 27 27 **22** 24 37 27

Table 2: Tour Measurements

Tour Number	Order of starting cities according to the scheme (1) tour	Index among all tours of levels 0_0	Tour length of level 0_0	Tour length of level 0i (starting to calculate)	Order of three-member tour cities (*level* 0_i)	Tour length of level 1_0
69	3 10 3	1	33633	35877	3 **22** 10 3	34479
70	3 43 3	1	33633	35254	3 **22** 43 3	34479
331	24 38 24	33	34151	35539 36239	24 **21** 38 24 24 **40** 38 24	34668 34151
332	24 43 24	33	34151	35131 35420 36138 35880	24 **1** 43 24 24 **8** 43 24 24 **22** 43 24 24 **40** 43 24	33879 33985 33879 34151
333	38 43 38	33	34151	36575	38 **22** 43 38	34510
394	22 24 22	46	34276	35881 35128 35561 36138 34483 36885 36138	22 **1** 24 22 22 **6** 24 22 22 **8** 24 22 22 **17** 24 22 22 **27** 24 22 22 **37** 24 22 22 **43** 24 22	33879 33955 33985 33955 33955 33955 33879
639	22 43 22	90	34479	36427 35137 35254 35818 36152 36138 37297	22 **0** 43 22 22 **1** 43 22 22 **3** 43 22 22 **9** 43 22 22 **20** 43 22 22 **24** 43 22 22 **38** 43 22	33991 33879 34479 33991 34094 33879 34510
854	10 22 10	156	34755	35877 35022	10 **3** 22 10 10 **28** 22 10	34479 34094
1127	12 38 12	258	35400	35659	12 **28** 38 12	34603

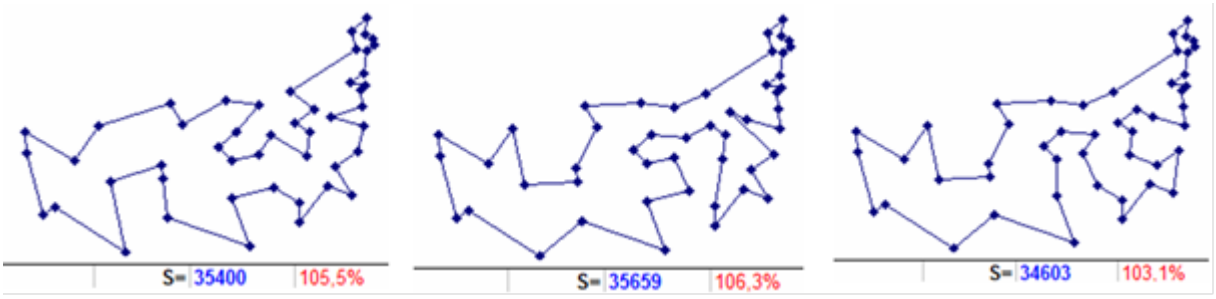


Figure 8. Routes from the starting cities 12 and 38 corresponding to the constructions according to the scheme (1): level 00 (left), level 03 (center), level 03 - 10 (right). Additionally, the length and deviation relative to the optimum are shown.

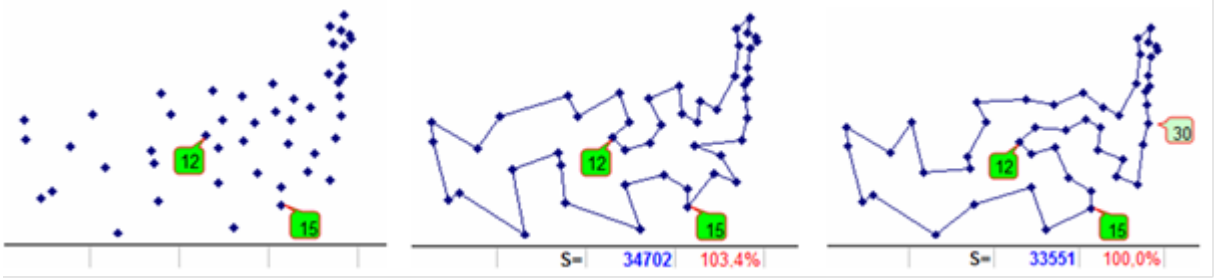


Figure 9. Cities 12 and 15 that make up the starting two-member tour (left). The route of level 00, corresponding to the constructions according to scheme (1), implies the inclusion of city 30 (in the center). Route level 00 - 10 (right), resulting from the inclusion of the city 30 (optimal). Additionally, the length and deviation relative to the optimum are shown.

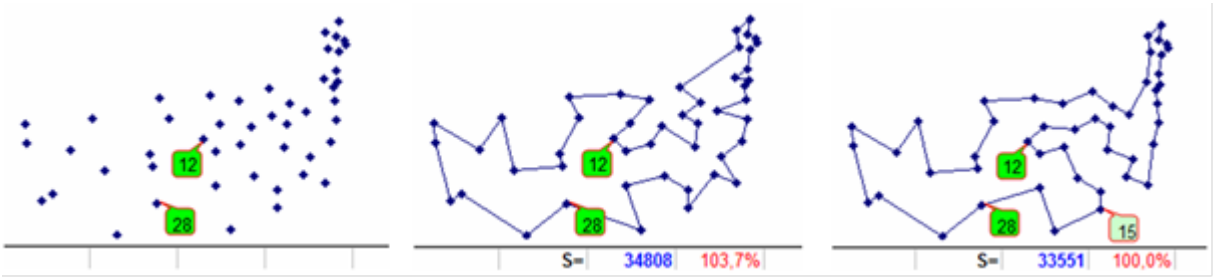


Figure 10. Cities 12 and 28 constituting the starting two-member tour (left). The route level 00, corresponding to the constructions according to scheme (1), implies the inclusion of the city 15 (in the center). Route level 00 - 10 (right), obtained as a result of the inclusion of the city 15 (optimal). Additionally, the length and deviation relative to the optimum are shown.

Samples of construction 10. It is necessary to separately distinguish the group of constructions where the optimal tour is obtained at the earliest stage (from the beginning of the calculations) when implementing the inclusion in the starting two-member tour of the third city, i.e., after calculations at the zero level.

Fig. 9 (left) shows an illustration of the starting conditions for the calculation of a two-member tour between cities 12 and 15. In the center of this figure, the shortest route (length 34702) is being built for those cities which have a zero level route (level 00). It is the only one and involves the inclusion in the two-member tour of the city 30. After the realization of this inclusion, we obtain the optimal route Fig. 9 (right).

Also, when implementing the inclusion of level 00, an optimal route is constructed from a two-member tour between cities 12 and 28. Fig. 10 (left) is an illustration of the starting calculation conditions for this two-member tour. In the center of the figure, the shortest (length 34808) zero level route (level 00) is shown. There are two such routes and, accordingly, two directions of construction according to the scheme (1). The first one, which implies the inclusion of city 2, is a dead end one for reaching the optimum. The second one, which implies the inclusion of the city 15, after its implementation immediately builds the optimal route Fig. 10 (right).

For more commonality, Fig. 11 shows 4 more routes of the zero level from binomial tours, but with a non-zero index (levels 0i, i $\neq$ 0: the first row in the center and to the right, the second row in the center and to the left). Each of the directions corresponding to them, after the implementation of the proper inclusion of the third city in the corresponding route of the two-member tour, also immediately leads to an optimal result (Fig. 11, second row rightmost). This happens again when constructing from the zero level, but only already when implementing the inclusion corresponding to a non-zero index (from the level 0i, i $\neq$ 0).

Unfortunately, in the absence of a general criterion for choosing the "correct" pairs of starting two-member tours, in order to identify all three-member tours, from which the optimal route is immediately constructed, 17296 possible options will have to be considered (in general, n * (n-1) * (n-2) / 6, where n is the dimension of the problem).

5. Features of finding the best route in tasks with a large dimension

In the light of the innovations outlined in the previous section, at the very beginning, there emerges a very large number of directions of calculations from each of which it is potentially possible to get the best route.

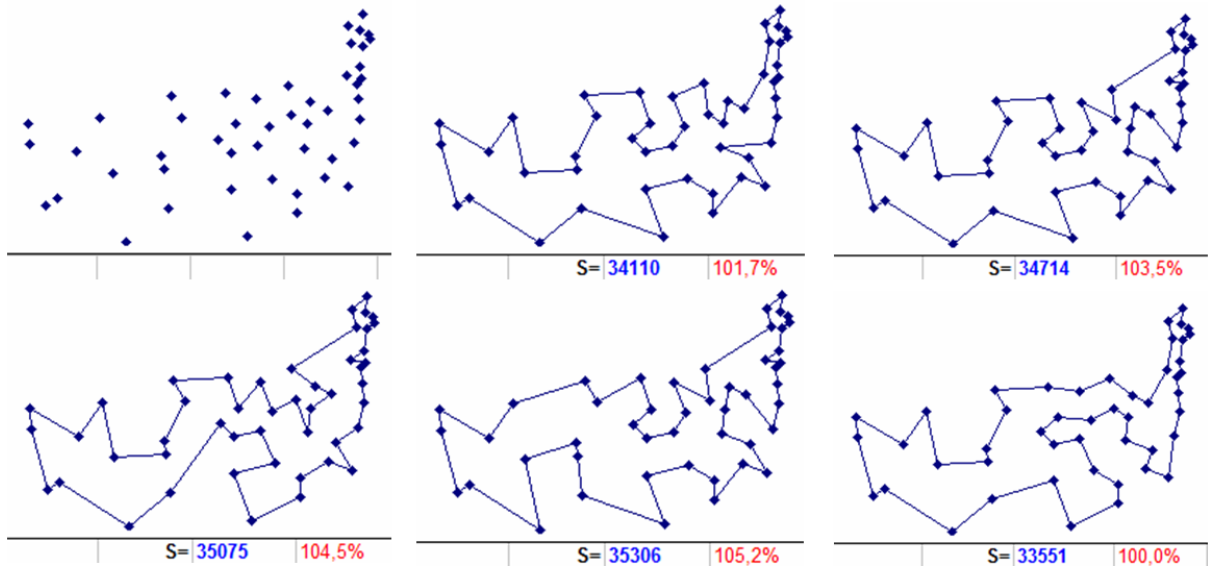


Figure 11. The mutual position of 48 cities (upper left). The optimal route is 33551 long (bottom right). The rest: routes of level 0i, directly transformed after the inclusion of the corresponding third city in the optimal route. Additionally, the length and deviation relative to the optimum are shown.

In such examples as ours (dimensionality of 48 cities) and even twice as large dimensionality, it is still possible to authorize the definition of not only the best route, but also the youngest level where it is reached (along with a full set of options for building it on it). This was shown in construction samples 10.

For examples with a dimension somewhat greater than 100 elements, such an approach (i.e., the calculation of "breadth") is already quite laborious. Therefore, we first denote a simplified version of the organization of calculations among the set of initial two-member tours. For each unique pair of source cities, determine the minimum route of level 00, and the corresponding city for inclusion in the two-member tour (if there are several, then one of them). Index all (n * (n-1) / 2) obtained routes in order of increasing length (we call these indices global). Limit the number of calculations from starting two-member tours (or groups of tours) only to those of them that have small global indices.

In addition, we will somewhat limit the number of calculations in scheme (5). One of its possible modifications can be formulated as the scheme (6) determined by the next paragraph.

If at the levels of J0, where J = 1,2,3 ..., only one route is observed (with index 0), then further constructions will only lead from the initial (J + 2) -member tour, i.e., according to the scheme (1).If there are several routes with index 0, then further constructions will lead not only from the initial (J + 2) -member tours for them, but also from all other (J + 2) -member tours that are initial for the level routes with corresponding non-zero indices, i.e., as in the scheme (4). At the same time, in order to further reduce the computations, as compared with scheme (5), we shall consider only separate directions of calculations from non-zero indices of the current level ("current-baseline"). These are the directions from which a chain of routes with zero indices of subsequent levels with decreasing lengths is built without branching. Moreover, the length at the end of the chain (before branching occurs), even if only one link actually turns out, should be less (or at least not more) than the length of the route with zero index of the "current-base" level.

Let us consider how this is implemented using the example of calculation, for a test with a dimension of 131 elements (Fig. 12, top row to the left) from the ASIC data set of the VLSI TSP collection (file xqf131.tsp) [20]. For convenience, we will use the numbering of the test elements (in this case, vertices or points) defined by a set of Cartesian coordinates, starting from 0 (and not from 1, as in the original source [20]).

Sample of construction 11 (from a two-member tour with the global index of 0). Among the routes of level 00, constructed from all possible combinations of initial two-member tours, the shortest length is 580.23 (Fig. 12 upper row in the center - for simplicity, we shall ignore digits after hundredths) and, accordingly, two of them have a global index of 0. The first one combines the vertices 28 66 and the second one has the vertices 28 62. Both assume that vertex 111 is included in the initial tour. Let us consider the constructions from the first one.

Thus, from the three-member tour (28 111 66 28) a minimum route is built (Fig. 12 upper row to the right) with a length of 579.99. It predetermines 3 variants of the inclusion vertices (25, 26 or 18) corresponding to the zero level index, and therefore along with them, according to scheme (6), other remaining vertices of the possible inclusion should be considered. Among all the directions, the inclusion of the vertex 34 (28 34 111 66 28) seems to be one of the most promising for further development. After it, a diminishing chain of lengths for complete routes is constructed with zero indexes of subsequent levels up to level 5, where there are three routes with zero index length 574.77 (Fig. 12 bottom row to the left) again. Therefore, the condition of the scheme (6) for the final length of the chain is less than the length of the route with the zero index (579.99) of the "current-base" level. Since then there are three inclusions (from zero level indices) into a seven-member tour (28 34 97 111 106 96 66 28), all remaining vertices with nonzero indices should also be considered. At this stage, there is a good potential for the development of directions at the inclusion vertices with zero indices (60 94 and 95). Consider the first of them - the inclusion of vertex 60 (28 34 97 111 106 96 60 66 28), from which a sequence of decreasing routes to level 8 is unambiguously constructed, where 36

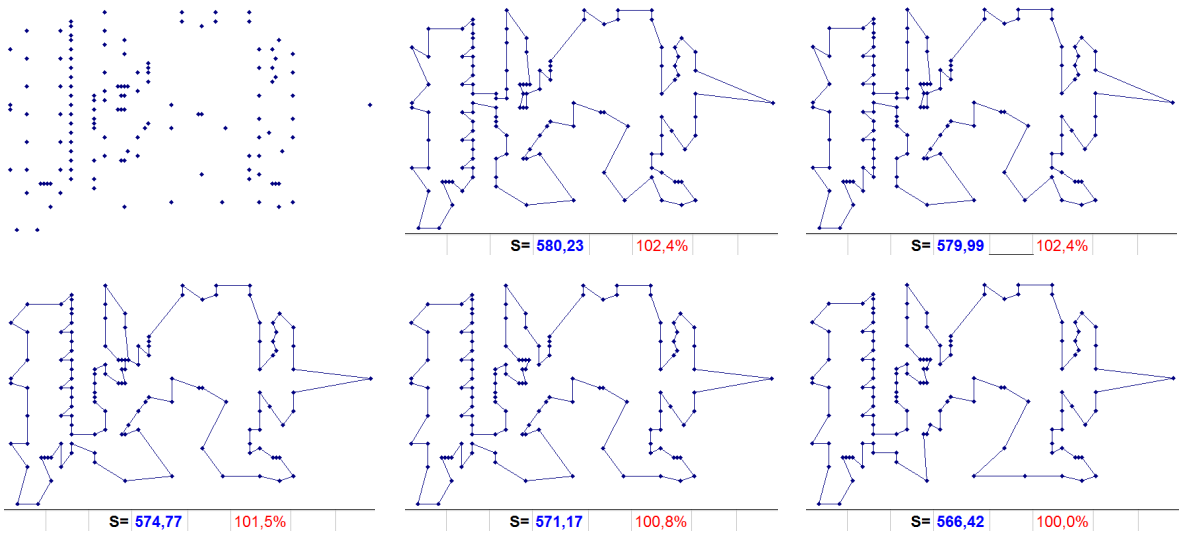


Figure 12. The mutual position of 131 points (top left). Record route length 566,42 (bottom right). The rest: a sequence of key intermediate routes. Additionally, the length and deviation relative to the record are shown.

directions (vertices) characterizing routes with a length of 571.17 with a zero index appear (Fig. 12 bottom row in the center). Any of them does not reduce the length of the route with a zero index on the next level. But from one of them - from the inclusion of vertex 93 in the ten-member tour (28 34 11 93 97 111 106 79 79 60 66 28), 44 more routes are formed with a zero index and with the same length. And here already (when constructing routes of the next level), a record 566.42 length is not improved (Fig. 12 bottom row to the right), from the inclusion of vertices corresponding to non-zero level indices. This happens when vertices 77, 87 (correspondence to index 4) or 13 (correspondence to index 9) are included between vertices 11 93 in tour 28 34 *11 93* 97 111 106 96 79 60 66 28.

Here is the record tour: 66 62 57 58 59 60 72 71 79 78 82 83 84 85 89 90 94 95 96 103 102 110 109 108 114 118 115 119 116 121 128 127 126 130 125 124 123 112 107 106 105 101 100 99 104 113 117 120 129 122 111 97 88 92 98 93 91 87 86 81 80 77 76 74 67 63 73 52 44 25 26 18 24 16 15 14 13 17 12 4 11 5 0 6 7 1 2 8 9 3 10 23 43 42 41 40 39 22 38 37 36 21 35 34 33 20 32 31 30 19 29 28 27 45 53 54 46 47 48 49 50 51 56 55 61 64 68 65 69 75 70 66.

It is achieved according to the scheme (6), at constructions from level 9 (00 – 163 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 9i, где i=4, 9), which is two levels earlier than at calculation according to scheme (5) from a two-member tour with the most distant cities in [15]. If the optimal route stated in [20] is calculated with an accuracy that is used in applied calculations for CNC (conventionally with “computer” accuracy), then we obtain the length 567.20, i.e. Our record route was shorter. In the integer metric used in [20], our route will be one unit longer.

Sample of construction 12. Let us check whether the path of obtaining a record result in certain areas will be shorter if we use scheme (1) for calculations from the same two-member tour with a global index of 0, as in construction 11.

And this direction is found. Let us consider the same two-member tour which unites the vertices of 28 66, but with the vertex of embedding 34, which in the framework of scheme (1) corresponds to the construction from level 084 with a route of 630.94 in length (Fig. 13 on the left).

From it, there are unambiguous constructions of improving routes with zero indices of subsequent levels up to the formation of routes of level 5. Where it appears after being included in the six-membered tour of vertex 11 (28 34 11 13 80 86 66 28), 33 routes with zero index that duplicate the route of level 40 (Fig. 13 in the center).

From it, there are unambiguous constructions of improving routes with zero indices of subsequent levels up to the formation of routes of level 5. Where appear, after being included in the six-membered tour of vertex 11 (28 34 11 13 80 86 66 28), 33 routes with zero index that duplicate the route of level 40 (Fig. 13 in the center).

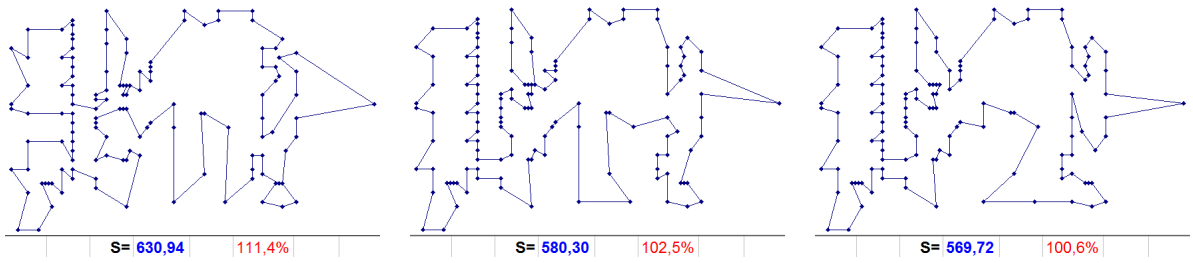


Figure 13. The sequence of key intermediate routes. Additionally, the length and deviation relative to the record are shown.

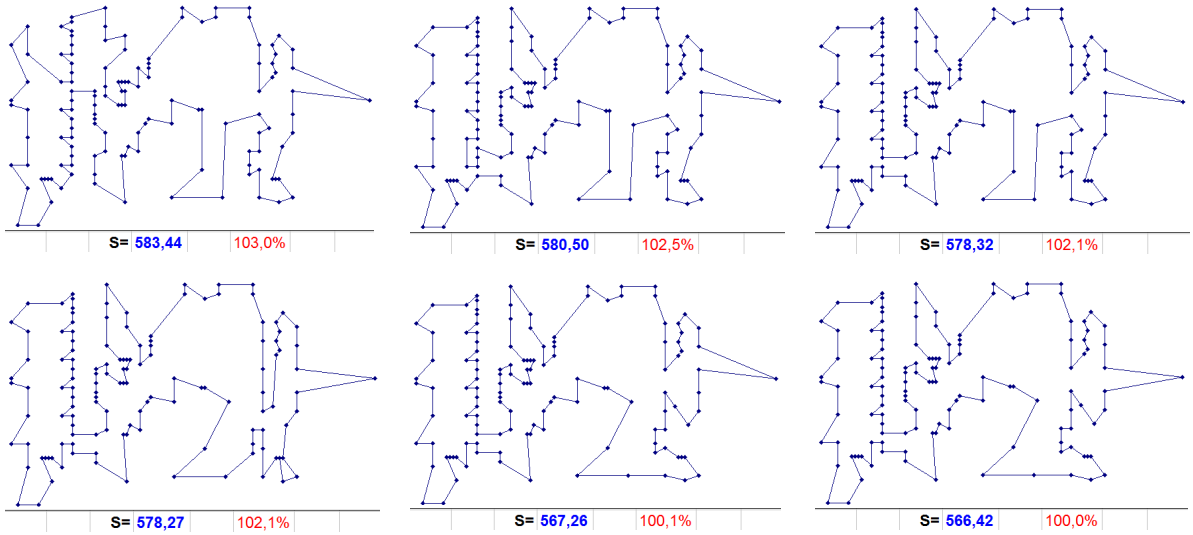


Figure 14. The sequence of key intermediate routes leading to the record. Additionally, the length and deviation relative to the record are shown.

From one of them (as a result of consecutive unambiguous inclusions in the seven-member tour 28 34 11 13 80 86 66 28 vertices 122 106 102), a chain of routes with zero indices of subsequent levels with a minimum length of the resulting route equal to 569.72 is built without branching (Fig. 13 on right). Moreover, there are two such routes and two new directions respectively from the inclusion of vertices 100 and 99 (28 34 11 13 80 86 122 100 106 102 66 28 - the first, and 28 34 11 13 80 86 122 99 106 102 66 28 - the second). Both directions, in order to reach the record route with a length of 566.42 (Fig. 12 bottom row to the right) assume that the 75 vertex is definitely turned on between 28 and 66, and at the next level 100 - alternative inclusions of the vertex 108 (between 102 and 106).

Sample of construction 13. The record route can be calculated at earlier levels by a repeated number of times and also according to a less laborious scheme (1), if not limited to the calculation, only from the global zero index.

So, for example, among the routes built from all possible combinations of the original two-member tours, the route from the two-member 6 6 6 6 level 00 with a length of 583.44 (Fig. 14 at the top left) has a global index of 16. It alone determines the inclusion of the 87 in the two-member tour.

Then, there is an unambiguous inclusion in the three-member tour of the vertex 54 (6 87 89 54 6), after which 9 routes appear with the index 0 level 2 with a length of 580.50 (Fig. 14 in the center above). Among the corresponding vertices of inclusion, only embedding vertices 4 and 112 reduce the route to the next level. Let us track the first of them (6 4 87 89 54 6).

From it, there is an unambiguous reduction in the length of the routes to the formation of tours of level 5, where 8 routes appear with the index 0, length 578.32 (Fig. 14 at the top right). The corresponding embedding in each of the tours of level 50 again multiplies routes with index 0 of the next level (an average of 8 routes without a change in length). And only embedding already in one of the many tours of level 60 leads to a decrease in the length of the route of level 70 to 578.27 (Fig. 14 at the bottom left).

This reduced route resulted from the inclusion of the vertex 88 in the eight-membered tour (6 0 11 4 87 88 97 89 54 6). The constructed route predetermines three possible embeddings in the tour 6 0 11 4 87 88 97 89 54 6 peaks 98 93 or 91.

Each of these inclusions gives a new reduced route with a length of 567.26 (Fig. 14 in the center below), and, accordingly, 3 initial tours with a unique inclusion of vertex 106. This inclusion in any of the 3 tours immediately gives a record route (Fig. 14 at the bottom right) , which repeatedly manifests itself in different directions when building at level 8. Thus, up to a record from the initial two-member tour 6 89 6, it was built on the levels 00 - 10 - 20 - 30 - 40 - 50 - 60 - 70 - 80 in accordance with the scheme (1).

The record route can be obtained even earlier - at level 70. With similar constructions according to scheme (1), from a two-member tour, 89 87 89 with a sequence number (lengths of level 00 routes, arranged in ascending order) are approximately midway between the same numbers for “sample build 11” and “sample build 13”.

6. Conclusion

A significant number of two-member tours (proportional to the square of the dimension of the problem), practically from each of which with varying degrees of difficulty can be built a record route, at least close to the optimal result, allow us to draw some preliminary conclusions based on the results of a set of calculations.

It can be argued with a high degree of probability that a significant layer of applied problems associated with the efficient operation of CNC equipment, equivalent in setting to a symmetric traveling salesman problem, can be solved using the method of step-by-step approach to the minimum route.

Moreover, many problems with a dimension of up to 50 elements, close to the present limit of the possibilities of exact methods, can be solved with obtaining an optimal result already at an early stage of calculations. Namely, when calculating at the level of possible combinations of two-member (formation and ranking of routes of zero level) or three-member (formation of routes of level 10) tours.

For problems of greater dimension, especially over 100 elements, it is advisable to reduce the calculations using one of the proposed schemes for organizing calculations (or its modifications). And as experiments show, you can use a more favorable criterion for stopping calculations in the current direction, for example, not threefold (as recommended), but double “repetition of the length of routes”. In most cases, this will only lead to a reduction in the directions of calculations leading to a record result, but on the whole, it will accelerate (n - fold) the calculation process.

References

[1] Nikolay Starostin, Konstantin Mironov: Algorithm Modification of the Level-by-Level Approximation to the Minimum Route, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 270-275, December 2017.

[2] N. D. Starostin The method of level-by-level exceptions for optimization of tool movements (in Russian) Collection of materials of the Second International Conference "Intelligent Technologies for Information Processing and Control", Ufa, 2014.

[3] N. D. Starostin. Improvement and detailing of the method of level-by-level exceptions for optimization of tool movements (in Russian) // Information Technologies and Systems [Electronic resource]: tr. Fourth Intern. sci. Conf., Bannoe, Russia, Feb. 25. - March 1, 2015, pp. 33-44

[4] I. I. Melamed The traveling salesman's problem. Exact methods (in Russian) II Melamed, SI Sergeev, I. Kh. Sigal // Avtomat. and telemeh. - 1989. - No. 10. - P. 3-29.

[5] V. V. Borodin Experimental study of the effectiveness of heuristic algorithms for solving the traveling salesman problem (in Russian) VV Borodin, SE Lovetskii, II Melamed, Yu. M. Plotinsky // Avtomat. and Telemeh., 1980, No. 11, 76-84.

[6] Starostin N. D. Finding the shortest path. Accelerated optimization of tool movements - Saarbrucken: LAP LAMBERT Akademic Publishing, 2015. - 89 p.

[7] N. D. Starostin Variants of calculation of the minimum route of processing (in Russian) Information technologies and systems [Electronic resource]: tr. Fifth International. sci. Conf., Bannoe, Russia, 24 February. - February 28, 2016, pp. 42-48.

[8] N. D. Starostin Basic variation of the algorithm of level-by-level approximation to the minimal route [Electronic resource, in Russian]: tr. The Sixth Intern. sci. , Bannoe, Russia, March 1 - 5, 2017 P.295-302.

[9] N. D. Starostin , DV Kurennov Level-by-level approach to the minimum route // International scientific and technical conference "Prom-Engineering" in 2017 was held May 16-19 in St. Petersburg.

[10] MP-TESTDATA - The TSPLIB Symmetric Traveling Salesman Problem Instances – URL: http://elib.zib.de/pub/mp-testdata/tsp/tsplib/tsp. (vizited on 08.02.2016).
[11] Lmatrix Laboratory Featured articles Solutions - URL: http://lmatrix.ru/news/problems (reference date: 01/05/2016)
[12] I. I. Melamed The traveling salesman's problem. Approximate algorithms (in Russian) / II Melamed, SI Sergeev, I. Kh. Sigal // Avtomat. and telemeh. - 1989. - No. 11. - P. 3-26
[13] Golden B. L., Skiscim C. C. Using simulated annealing to solve routing and location problems // Nav. Res. Log. Quart. 1986. V. 33. № 2. P. 261-279
[14] National TSP – URL: http://www.math.uwaterloo.ca/tsp/world/countries.html (vizited on: 05.01.2017
[15] Yu. A. Kochetov, Computational possibilities of local search in combinatorial optimization, Zh. Vychisl. Math. and Math. fiz. - 2008 - Vol. 48, No. 5 - P. 788-807.
[16] An Interview with Vasek Chvatal. – URL: https://www.cs.rutgers.edu/~mcgrew/Explorer/1.2/#Farmers (visited on 16.08.2017)
[17] Kochetov Yu., Mladenovich N., Hansen P. Local Search with Alternating Neighborhoods // Discrete Analysis and Operations Research. Series 2. 2003. Volume 10, No. 1. P.11-43.
[18] Metaheuristics: From Design to Implementation. El-Ghazali Talbi. John Wiley & Sons, 2009, - 624 p
[19] R. T. Murzakaev, V. S. Sholov, A. V. Burylov. Application of Metaheuristic Algorithms for minimizing the path of a cutting tool (in Russian) Transactions of PNIPU University on Electrical Engineering, IT, and Control. 2015. №14. C.123-133
[20] VLSI TSP Collection – URL: http://www.math.uwaterloo.ca/tsp/vlsi/index.html (дата обращения: 31.08.2017)
[21] MP-TESTDATA - The TSPLIB Symmetric Traveling Salesman Problem Instances – URL: https://people.sc.fsu.edu/~jburkardt/datasets/tsp/tsp.html (дата обращения: 06.02.2018)
[22] In Pursuit of the Traveling Salesman : Mathematics at the Limits ofComputation. Author(s), Cook, William J. Publication, Princeton, NJ : Princeton University Press, 2011. - 245 p.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 282-291 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Over-The-Air Testing of Automotive Antennas and Wireless Links in The Installed State on The Basis of LTE Downlink Communication Parameters

Philipp Berlt*, Lisa Jäger, Andreas Schwind, Frank Wollenschläger, Christian Bornkessel, Matthias A. Hein

Thuringian Centre of Innovation in Mobility, RF and Microwave Research Laboratory, Technische Universität Ilmenau, 98693 Ilmenau, Germany

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 28 January, 2019
Online : 20 February, 2019

Keywords:
Antenna measurement
LTE
Over-the-air

ABSTRACT

Modern automobiles have been turning more and more into wireless sensor and communication networks. Accordingly, the number of radio systems is steadily increasing. Due to strict safety requirements, these radio systems need to be tested extensively for functionality and reliability, especially under poor radio channel conditions. Beside the large electrical size of cars at frequencies of mobile communication services, access to the antennas imposes challenges for testing, due to an increasingly high integration of the antennas with frontends and digital signal processing. This paper proposes an over-the-air testing procedure for automotive radio systems on the basis of the wireless communication standard Long Term Evolution (LTE). A method to derive the radiation patterns of automotive antennas from reference signals in the LTE downlink scheme without requiring any access to the analogue RF feed point is proposed. A comparison of the LTE approach with the usual antenna measurement techniques shows good agreement. As a logical step from the antenna towards a complete wireless link, a concept of spatially distributed channel emulation on the basis of Software Defined Radio (SDR) modules is proposed, aiming at the emulation of essential multipath features of the wireless channel. RF measurements of the channel transfer function as well as by over-the-air end-to-end tests prove this approach to be a cost-efficient alternative to commercial channel emulators.

1. Introduction

This paper is an extension of work originally presented at the 2018 IEEE 87th Vehicular Technology Conference [1]. The paper deals with automotive antenna radiation pattern measurements on the one hand, and with end-to-end communication testing on the other hand. Both subjects are closely related, as the antenna performance has an essential impact on the overall system performance. Moreover, wireless signal transmission is strongly affected by the radio channel conditions. As a consequence, end-to-end testing of the entire communication link, including application antennas as well as radio channel conditions, is essential in order to ensure reliable functionality in real applications.

Connected and highly automated driving is a key to future mobility, especially in terms of increasing safety and improving traffic flow. Vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) communication, generalized as vehicle-to-everything (V2X), has been of great interest in research and development and pushed forward by industry organizations like the 5G Automotive Association (5GAA), Car to Car Communications Consortium (C2CCC), or the Virtual Drive Test Alliance (VDT Alliance) [2][3][4]. There has been a controversial discussion about optimal technologies for the implementation of V2X networks. Beside WLAN-related networks like IEEE 802.11p, derivatives of the LTE standard are under consideration [5]. This paper focuses on the latter due to its greater variability in terms of cellular and ad-hoc networking, and in terms of its applications in mobile communications, data transmission, and V2X communications. When moving towards higher levels of driving automation, the number of sensor nodes and wireless systems in modern automobiles will increase significantly, e.g., in order to enable cooperative perception or manoeuvring. Satellite-based navigation, automotive radar sensing, but also mobile communication systems will provide altogether the basis for smart cooperative connected cars. Due to the increasing amount of wireless air interfaces, the installation space for antennas is seriously limited and, thus, the future of antennas lies in integrated modules, where radiating elements, frontends, and signal processing units are merged. As a consequence, access to the analogue RF antenna feed cannot be taken for granted any more.

*Corresponding Author: Philipp Berlt, E-Mail: philipp.berlt@tu-ilmenau.de

Consequently, conventional antenna measurement techniques, e.g. far-field measurements or nearfield measurements with subsequent nearfield-to-far-field transformation [6] [7], seem unsuitable for such integrated systems. Hence, alternative test procedures need to be identified and investigated. In this paper, we propose a concept that allows for the derivation of automotive antenna gain patterns in their installed state on the basis of LTE communication parameters, monitored by a commercial communication modem with USB drive. Instead of a wired RF connection to the antenna feed, a serial connection to a laptop computer is required, in order to retrieve information about the signal quality from the modem.

In addition to the radiation patterns of the antennas, the overall performance of automotive wireless communication systems needs to be tested, in order to ensure functionality even under poor mobile channel conditions. Drive tests have been the procedure-of-choice for the verification and validation of automotive systems. However, these tests are expensive and time consuming and, due to changing conditions in a real-world environment, reproducibility cannot be maintained. Therefore, time- and cost-efficient test procedures under reproducible laboratory conditions need to be investigated, leading to the concept of virtual drive tests. One approach for testing mobile devices under realistic radio channel conditions is the electromagnetic wave field synthesis. However, this approach is not feasible for electrically large objects like cars, as the size of the test zone is strictly limited by the available number of illumination antennas and channel emulators [7]. Therefore, in this paper, we follow a cluster-based concept for spatially distributed channel emulation, which has been implemented using commercially available Software Defined Radio (SDR) modules like the Universal Software Radio Peripheral (USRP) [8]. The functionality of the channel emulator has been demonstrated by end-to-end LTE measurements as well as by RF measurements of the emulated transfer function. All measurements were conducted in our multipurpose test facility virtual road – simulation and test area (VISTA).

2. Automotive test facility "Virtual Road Simulation and Test Area"

The unique automotive antenna test facility VISTA was installed at the Technische Universität Ilmenau under the roof of the Thuringian centre of innovation in mobility [9]. It is a multi-purpose test and measurement facility, including over-the-air testing and channel emulation for automotive wireless systems, three-dimensional radiation pattern measurements of antennas in their installed state, and electromagnetic compatibility testing. The facility comprises a shielded semi-anechoic chamber, a spherical nearfield antenna measurement system, a roller dynamometer integrated into an EMC-compatible turntable, and powerful software tools for the digital post-processing of the measured raw data. The walls and the ceiling of the chamber are covered with 60-inches pyramidal absorbers, designed for a lower frequency limit of 70 MHz. The metallic floor may be left blank, e.g. for radiated emission measurements, or covered with absorbers, e.g. for antenna measurements. The antenna measurement system is a spherical nearfield system, consisting of a multi-probe antenna arch for rapid electronically switched elevation scanning, a turntable for mechanical scanning along the azimuth, and a vehicle lift for elevating the car-under-test to the centre of the measurement arch at a height of 2.30 m above ground, as depicted in Figure 1. The left-hand semi-arch is equipped with 111 antennas covering the elevation range from 0° to 20° below horizon in 1° steps, for the frequency range from 400 MHz to 6000 MHz. The right-hand semi-arch contains 22 antennas from 2.5° to 17.5° below horizon with a spacing of 5°, for the frequency range from 70 MHz to 400 MHz. The resulting far-field radiation patterns are obtained by applying nearfield to far-field transformation algorithms. A more detailed description of VISTA can be found in [10].

3. Derivation of automotive antenna gain patterns from LTE downlink communication parameters

In this section, the procedure to derive gain patterns of automotive antennas in the installed state from LTE downlink parameters is described. In contrast to conventional antenna measurement techniques, the approach benefits from the fact that no RF cable connection to the antenna-under-test is required, rather than a serial connection to the LTE modem, which is interrogated by the connected laptop.

3.1. LTE hardware and transmission parameters and measurement setup

The commercially available communication tester CMW500 (Rohde&Schwarz) was used to establish the LTE link between an emulated base station and the LTE user equipment in VISTA [11], according to the setup outlined in Figure 2. This instrumentation allows for a full control of downlink and uplink parameters in terms of bandwidth, power levels, and resource allocation. Single-input single-output (SISO) as well as multiple-input multiple-output (MIMO) modes are supported for up to two receive and transmit antennas (2x2 MIMO). The antenna measurements described in this paper were conducted in SISO mode. The mobile terminal was a Huawei modem of type E3276s-150 (LTE category 4), connected to a laptop with a universal serial bus (USB). External monopole antennas were installed on the roof of a mid-sized passenger car mock-up and connected to the modem (see Figures 3 and 4). The modem includes one antenna for receive only, and a second antenna for reception and uplink signal transmission. Several power-related parameters, such as signal-to-interference-and-noise ratio (SINR), received signal strength indicator (RSSI), and reference signal received power (RSRP), are provided by the modem.

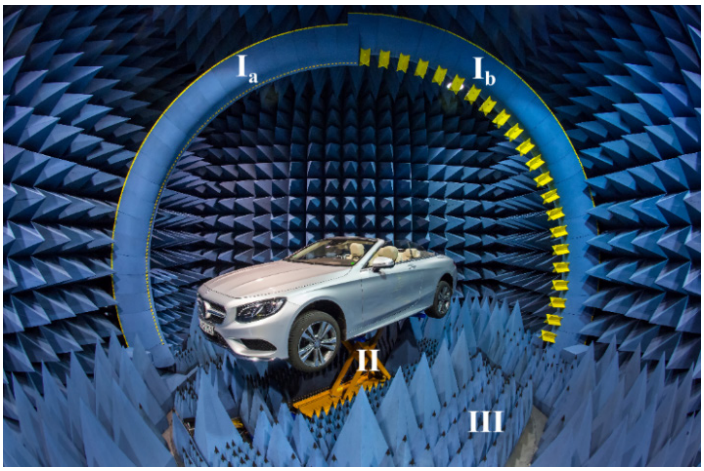


Figure 1: Photograph of VISTA with a car installed for automotive antenna measurement: Multi-probe antenna arch for the frequency range from 400 MHz up to 6 GHz (Ia) and from 70 MHz to 400 MHz (Ib). II: A customized vehicle lift elevates the car-under-test to the centre of the measurement arch 2.30 m above ground. III: The EMC-compatible turntable is used for azimuthal rotation with an accuracy of 0.1°.

RSSI is a measure of the total power received across the selected bandwidth. Its value depends on interference, noise, and channel traffic. In contrast, RSRP quantifies the average received power of dedicated reference signals (RS) in the LTE downlink. The reference signals are used for downlink channel estimation and are transmitted with constant power, independent of channel traffic [12]. For this reason, RSRP was selected as the parameter optimally suited for antenna gain measurements.

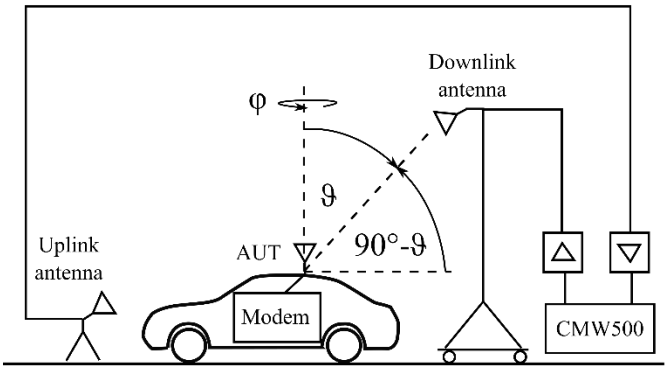


Figure 2: Sketch of the RSRP-based measurement setup. The downlink horn antenna is installed on a height-adjustable antenna mast. Depending on the elevation angle, 90° - ϑ, the height may be set between 1 m and 4 m. The distance between the downlink transmit antenna and the centre of rotation is 7 m. An additional uplink antenna is installed near to the car, in order to provide the feedback channel from the modem to the communication tester. The implemented setup is discussed below and can be seen in Figure 7.

Figure 3: Photograph of the antenna-under-test, installed in VISTA. The AUT (indicated by the black circle) is connected to the Tx/Rx path of the UE, while the second Rx path is matched with a 50 Ω load.

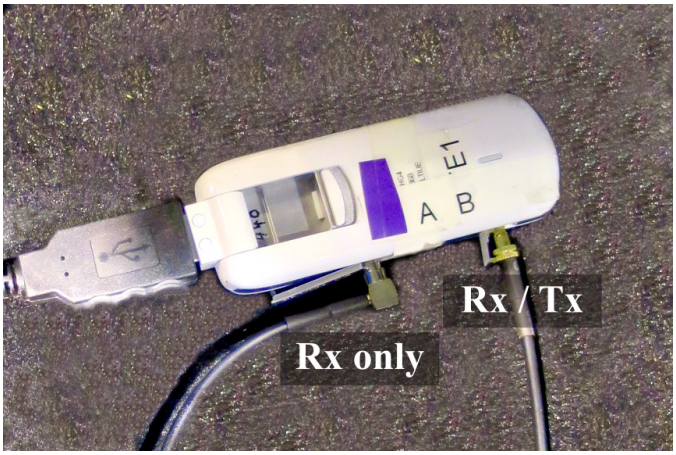


Figure 4: Photograph of the USB drive LTE modem, connected to external monopole antennas on the roof of the mock-up of a mid-size passenger car (see Figure 3). The AUT is connected to the Tx/Rx path of the UE, while the second Rx path is matched with a 50 Ω load.

In order to conduct reliable measurements with the user equipment, the input/output power relation of the UE needs to be calibrated. The modem was characterized in terms of linearity and dynamic range in a wired measurement setup as depicted in Figure 5. Figure 6 displays the results for the I/O relation of the modem.

The downlink transmit power, given in (dBm/15 kHz), was reduced stepwise by 0.25 dB and the received power was detected by the user equipment. The measurements were performed at a carrier frequency of 2655 MHz with 1.4 MHz channel bandwidth, which is the smallest bandwidth supported by the LTE standard. The bandwidth is a measure of the frequency resolution in the RSRP measurement. As expected, the RSRP follows the power

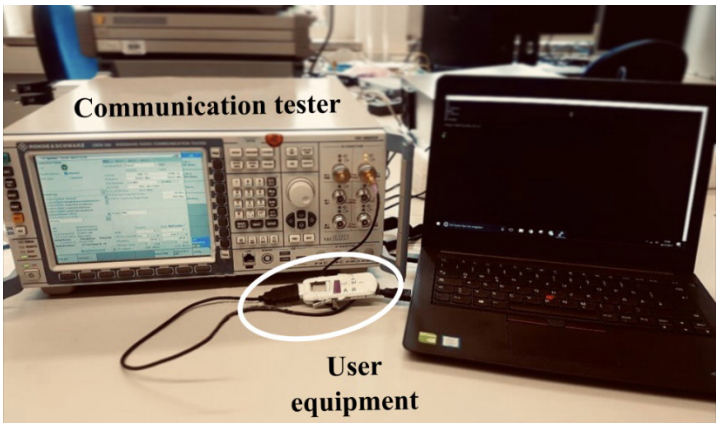


Figure 5: Setup for the measurement of the I/O relation of the LTE modem (right-hand side) with the communication tester (left-hand side). The downlink transmit power is detected by the user equipment.

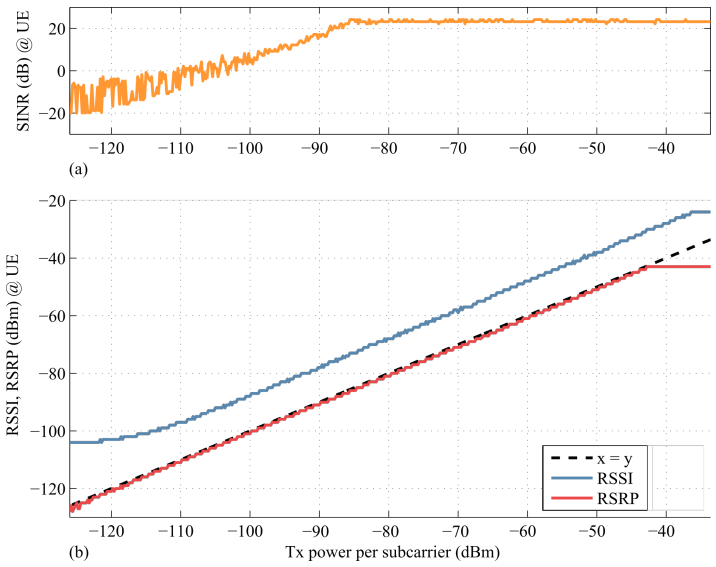


Figure 6: Typical transmission curves measured with the LTE modem. (a): The SINR (orange curve) increases linearly with the input power above the noise floor of about -100 dBm, and reaches a saturation level of 23 dB above -85 dBm. (b): The RSRP (red curve) follows the configured input power linearly for values less than -43 dBm, as indicated by the dashed black curve. The RSSI (blue curve) reaches saturation at approximately -110 dBm input power.

values set by the CMW500 very well, as it is a measure of the average power per RS resource element. Enabling detection of power levels less than -120 dBm, the RSRP offers high sensitivity, whereas the RSSI reaches saturation at -110 dBm input power. The measured data have a quantisation of 1 dB, set by the UE.

3.2. Radiation patterns derived from RSRP measurements

Antenna pattern measurement setup

A LTE link was established in VISTA between the communication tester and the antenna-under-test. The downlink signals were transmitted with a horn antenna (Schwarzbeck BBHA 9120 A) with a gain of 9.3 dBi, which was installed on a height-adjustable antenna mast (see Figure 2), in order to achieve various elevation angles-of-arrival. A second antenna was installed near the car, in order to receive the uplink signals transmitted by the user equipment. The turntable with the car was rotated stepwise, and the RSRP was evaluated as a function of the azimuth angle-of-arrival. The implementation is visualized in Figure 7. The distance between the downlink antenna and the AUT was $R_0 = 7$ m. Taking into account the height of the car and the antenna mast, elevation angles up to 15° were adjusted. The gain of the receive antenna G_{Rx} (AUT), in decibels, can be calculated from the RSRP according to [13]:

$$G_{Rx} = RSRP - P_{Tx} - G_{Tx} + L_{Cbl} + PL \quad (1)$$

P_{Tx} denotes the transmit power in dBm per 15 kHz, G_{Tx} denotes the gain of the horn antenna (downlink antenna) in dBi, L_{Cbl} and PL are cable losses and the free-space path loss in dB, respectively. For given wavelength λ and locations of transmit and receive antennas, the path loss is computed at each turntable angle ψ:

$$PL(\psi) = 20 \cdot \log_{10}\left(\frac{4\pi}{\lambda} \cdot \left| \tilde{R}_z(\psi) \cdot \vec{p}_{Rx,0} - \vec{p}_{Tx} \right|\right) \quad (2)$$

Chosen the center of rotation of the turntable as the origin of the coordinate system, $R_z(\psi)$ denotes the rotation matrix around the z-axis, p_{Tx} is the position of the transmit antenna, and $p_{Rx,0}$ denotes the initial position of the AUT at turntable position $\psi = 0°$. The location of the receive antenna is derived from phase center calculations, as further described in [14].

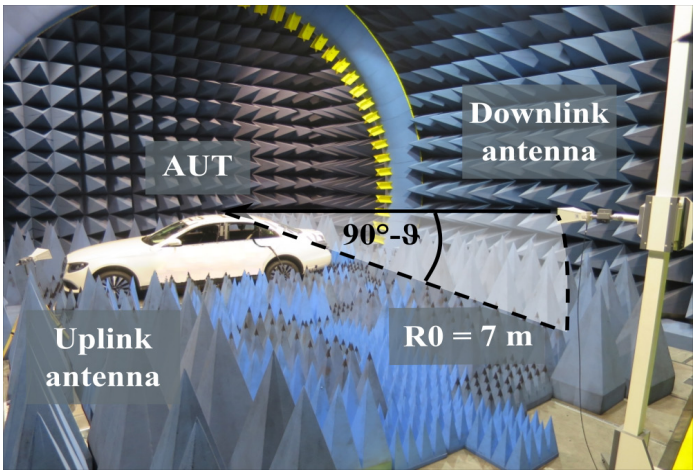


Figure 7: Photograph of the setup for the RSRP and VNA measurements. In order to suppress reflections from the metallic ground, the path between transmit antenna and antenna-under-test was covered with absorbers.

Comparison between RSRP approach and direct power transmission measurements with a network analyser

In order to validate the RSRP measurements, the gain of the AUT was additionally derived from calibrated power transmission measurements with a vector network analyser. The communication tester was replaced by the VNA, and the uplink antenna became obsolete. The positions of the transmit antenna and the AUT remained the same as in the previously described setup. The transmission factor $|S_{21}|_{dB} = 10 \cdot \log_{10}(|S_{21}|^2)$ was recorded at each angular step for a full 360° rotation of the car-under-test and, according to (3), the receive antenna gain was calculated from the measured transmission coefficient.

$$G_{Rx} = |S_{21}|_{dB} - G_{Tx} + L_{Cbl} + PL \quad (3)$$

Figure 8 displays horizontal cuts of the vertically polarised radiation pattern of the AUT at a co-elevation of $\vartheta = 10°$ and at a frequency of f = 2655 MHz, derived from both setups. A more detailed comparison, in terms of mean and variance of the error, is depicted in Figure 9 for a range of elevation angles. While the mean value indicates a static offset between the two datasets, the variance of the deviation reflects the similarity between the shapes of the patterns since it ignores a possible offset in the data. In order to express the similarity of two patterns by a single parameter, the cosine similarity can be calculated from (3) in analogy to antenna correlation [15].

$$\rho = \frac{\sum_m \sum_n G_1(J_m, \varphi_n) \cdot G_2(J_m, \varphi_n) \cdot \sin(J_m)}{\sqrt{\sum_m \sum_n G_1^2(J_m, \varphi_n) \cdot \sin(J_m)} \cdot \sqrt{\sum_m \sum_n G_2^2(J_m, \varphi_n) \cdot \sin(J_m)}} \quad (4)$$

G_1 and G_2 denote the gain patterns #1 and #2 on a linear scale. The weighting factor $\sin(\vartheta_m)$ compensates for the irregular spacing of the samples across the measurement sphere, given by ϑ_m and φ_n. According to (4), ρ can take values between 0 and 1, where a value of 0 implies orthogonal patterns and a value of 1 implies identical patterns, except for a scaling factor. Both datasets show good agreement regarding the angle-dependent shape of the radiation pattern, indicated by the low variance of the error (≤ 1 dB) and a high cosine similarity above 98%.

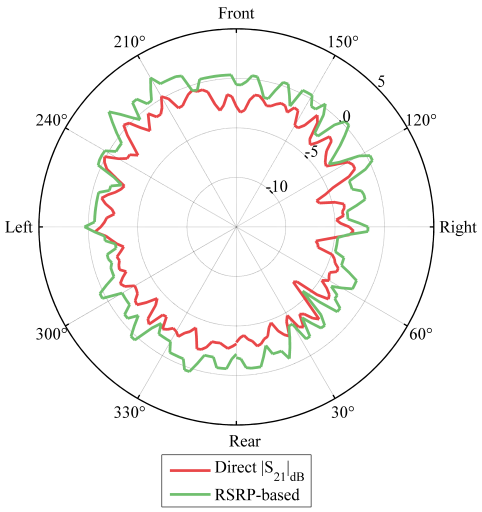


Figure 8: Comparison of the azimuth gain pattern measured via network analyser (red curve, $G_{max} = 0.4$ dBi) and RSRP data (green curve, $G_{max} = 2.4$ dBi) for an elevation angle of 10°.

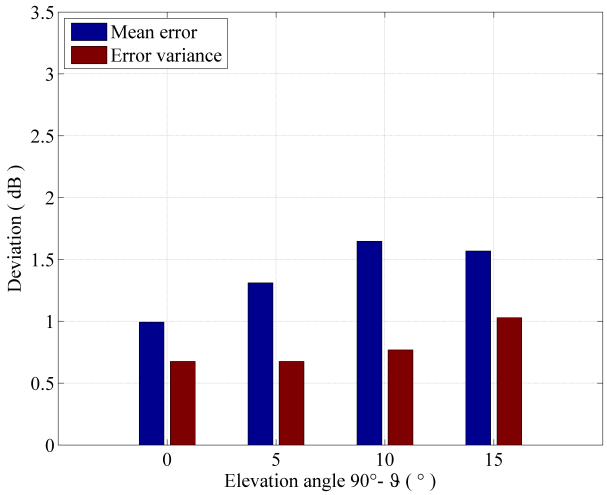


Figure 9: Bar diagram of the mean value (blue) and the variance (red) of the measurement deviation (in dB) between the direct power transmission measurement and the RSRP-based approach.

The RSRP approach tends to yield slightly higher gain values than the direct power transmission approach. Higher signal levels might be explained by measurement uncertainties during calibration and by the modem itself. Path loss and transmit antenna gain from (1) and (3) can be excluded as a reason for uncertainty between both datasets since kind and position of the antennas remained the same in the measurements and the same values were used in data analysis.

Comparison between RSRP approach and far-field pattern obtained from nearfield to far-field transformation

In addition to the RSRP and VNA measurements and the proof of their consistency, the AUT was measured with the spherical nearfield system in VISTA. For this purpose, the car mock-up was elevated, in order to place the antenna close to the phase centre of the measurement arch, as indicated by Figure 1. The far-field radiation pattern was derived from a nearfield to far-field transformation. The three-dimensional radiation pattern for vertical polarisation is depicted in Figure 10.

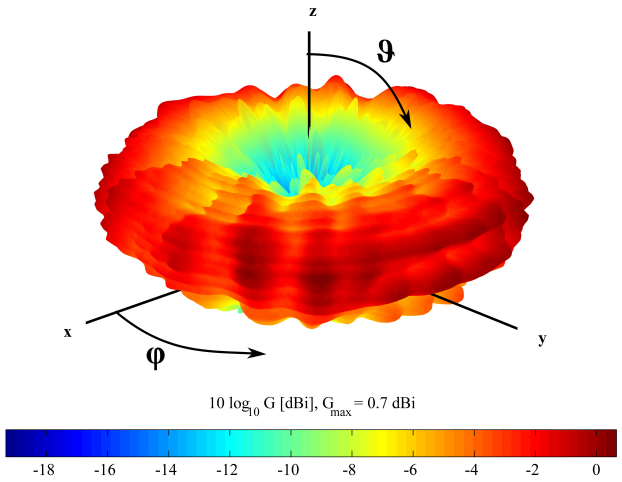


Figure 10: Three-dimensional realised-gain pattern derived from a nearfield to far-field transformation of the AUT for vertical polarisation at a frequency of 2655 MHz. The maximum gain value is 0.7 dBi.

The coordinate system was oriented as usual for automotive antenna measurements: the y-axis points along driving direction, and the z-axis to the zenith. The maximum realised gain was 0.7 dBi.

Since the vehicle was lifted to the centre of the measurement arch, a comparative measurement with the network analyser was conducted, in order to ensure comparability between the two setups with the car located on ground level and lifted, respectively. Due to the limited height of the antenna mast, just a single cut of the pattern for an elevation of 0° could be recorded with the direct power transmission method in the elevated state. Figure 11 shows that the two data sets compare favourably. The comparison between the RSRP and the far-field patterns is displayed in Figure 12, depicting the horizontal cuts of the gain patterns at an elevation of 10°. Figure 13 analyses the deviation between the data sets for the elevation angles considered. The patterns display very similar shapes, as quantified by a cosine similarity of 96 %.

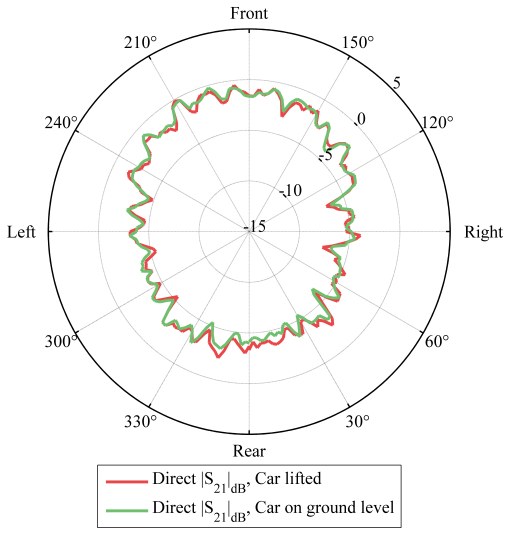


Figure 11: Comparison of the measurement setups in elevated state (red curve) and on ground level (green curve). The curves agree at all angles within 1.1 dB.

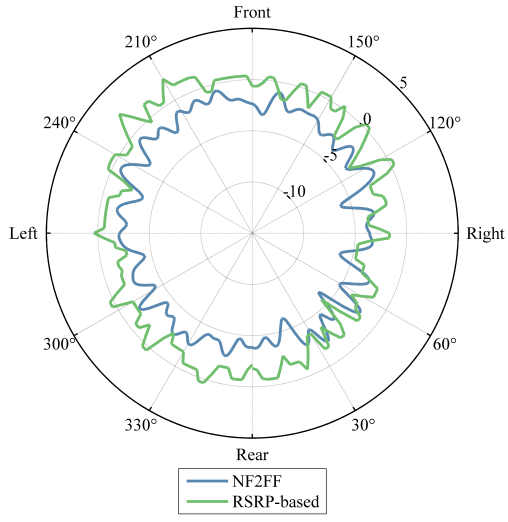


Figure 12: Comparison of the horizontal cuts at an elevation angle 10° of the gain patterns derived from the RSRP approach (green curve, G_{max} = 0.4 dBi) with the dataset obtained from nearfield to far-field transformation (blue curve, G_{max} = 2.4 dBi).

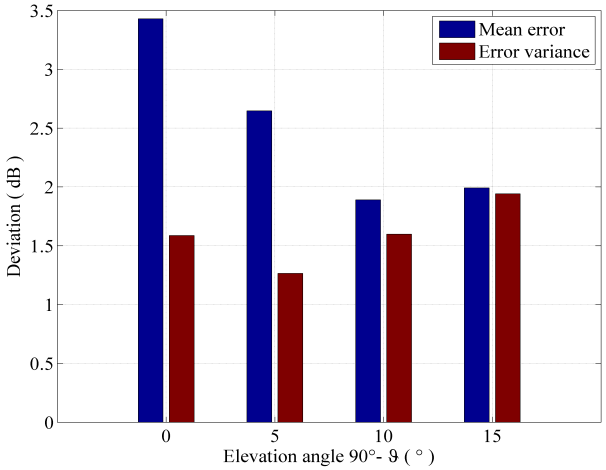


Figure 13: Bar diagram of the mean value (blue) and the variance (red) of the measurement error in dB between the RSRP nearfield pattern measurements and the nearfield to far-field transformation.

Summary of LTE antenna pattern measurements

Three types of measurements were performed, in order to derive horizontal cuts of the gain pattern of an automotive antenna: direct power transmission, nearfield measurement with subsequent nearfield to far-field transformation, and LTE-based RSRP measurement. The patterns derived from the three data sets compared very well in their shape. The small deviations observed between the absolute gain values of the three data sets may have been caused, in addition to the usual measurement uncertainties, by different post-processing of the measured signals and through reception of scattered signals by the user equipment within the car, partially bypassing the external antenna path. The user equipment was utilised as a measurement instrument, in order to evaluate downlink-specific reference signals and thus to obtain information about the received signal strength. It could be shown that the gain of the antenna can be derived from measured RSRP values without requiring access to the RF feed. Knowledge of the antenna gain is sufficient for many applications. The retrieval of phase information, following approaches like [16-18] will be in the focus of future work.

4. Over-the-air end-to-end testing of LTE links in a virtual electromagnetic environment

The overall wireless link performance of a device-under-test (user equipment) is affected by the wireless propagation channel, the antenna patterns involved in downlink and uplink communication, and the digital signal processing at both sides of the link. In order to exploit channel capacity more efficiently, multiple antennas may be installed on transmit and receive side (MIMO).

MIMO systems allow for distinctly higher data rates under favourable radio channel conditions, deploying multiple data streams by spatial multiplexing; in case of poor signal-to-noise ratio, transmission may be more robust against distortions through transmit diversity [19]. This section addresses end-to-end testing of LTE-based automotive wireless communication systems, identifying the realized data throughput as the figure-of-merit, as this describes the quality-of-service experienced by the end user best. The testing strategy follows a deterministic approach, based on the emulation of basic parameters of the wireless channel like propagation delay and Doppler shift. In contrast to stochastic fading models, physical propagation parameters are mapped deterministically onto a number of N illumination antennas. Furthermore, LTE-specific features like link adaptation and their effect on the link quality are considered.

4.1. Emulation of basic features of a wireless multipath channel

Mobile communications are characterized by multipath propagation. The signal received at the user equipment results from a superposition of many multipath components, including line-of-sight (LOS) and non-line-of-sight (NLOS) components. The latter result from reflections, scattering, refraction, or diffraction, and can be characterized by parameters like propagation delay, Doppler shift, amplitude variation, and phase shift. Additionally, spatial properties of the channel, like angles-of-arrival and angles-of-departure, need to be taken into account, especially when MIMO systems are considered. A time- and cost-efficient test procedure is favoured, reducing the amount of required hardware and software, but still projecting relevant real-world features onto the test scenario. Complexity may be reduced by aggregating several multipath components with similar properties into so-called clusters, either in the spatial domain or by joint evaluation of the angular direction-delay domains [20] [21].

In this study, channel emulation has been applied to the downlink signal by employing a number of Software-Defined Radio (SDR) modules (Universal Software Radio Peripherals, USRP) of type NI USRP 2954-R. For this purpose, an emulator previously developed on the basis of a PXI computer platform, developed by National Instruments, was adapted for the application of radio channel emulation synchronised between several USRP devices [22] [23]. The emulator architecture is divided into parameter generation on a host computer, and digital signal processing on the internal FPGA of the USRP. The USRP is controlled via a PCIe connection to the host PC, which allows for fast exchange of data and control traffic. As shown in Figure 14, each SDR module provides two RF channels, including digital down conversion (DDC) and digital up conversion (DUC), as well as the application of propagation parameters to the baseband input signal. The discrete impulse response of one specific path n of the distributed emulator can then be described by

$$h_n(\tau, t) = a_n \cdot e^{j(2\pi f_{D,n} t - \varphi_{0,n})} \cdot \delta(\tau - \tau_n) \quad (5)$$

with the Dirac delta function $\delta(\tau)$ [24], the path coefficient a_n, the static phase shift $\varphi_{0,n}$, the delay τ_n, and the Doppler shift $f_{D,n}$. Different directions-of-arrival can be emulated by a proper geometrical arrangement of the illumination antennas, placed around the device-under-test at a distance R.

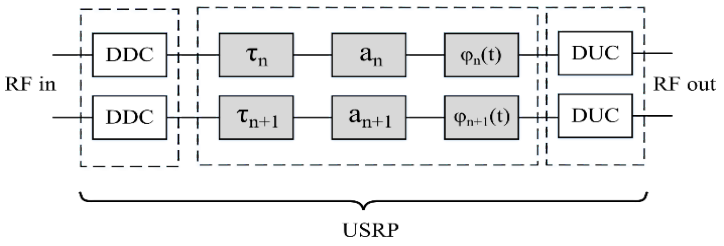


Figure 14: Simplified block diagram of RF signal processing implemented in the USRP. The analogue input signal RF in is converted via digital down conversion (DDC) to the digital baseband, where propagation delay τ_n, attenuation a_n, and a phase shift φ_n are applied. The phase shift may be time-dependent, in order to generate Doppler shifts. The modified RF signal RF out is then generated by digital up conversion (DUC).

According to (6), the resulting emulated channel impulse response is the superposition of time-delayed waves (wave number $k = 2\pi/\lambda$) impinging on the receive antenna, including the effects of antenna patterns and free-space propagation.

$$h(\tau,t) = \sum_{n=1}^{N} h_n(\tau,t) \times e_{Rx}(\Omega_{A,n}) \times e_{Tx,n}(\Omega_{D,n}) \times \frac{\lambda \times e^{-jkR}}{4\pi R} \qquad (6)$$

The quantities involved are the complex-valued transmit and receive polarimetric antenna element patterns e_{Tx} and e_{Rx}, corresponding to direction-of-departure $\Omega_{D,n}$ and direction-of-arrival $\Omega_{A,n}$, respectively. The proof-of-functionality of the emulator was achieved by measurements of the channel transfer function H(f) for given values of delay, Doppler frequency, and the amplitude ratio of multipath components, with transmitter and receiver connected by cables. The simple, yet very relevant two-path model was chosen as a reference scenario, in order to generate a transfer function with well-known characteristics suitable for comparison with analytical results. The LTE test signal was generated by an arbitrary-waveform generator and distributed to the USRP with a power splitter, as depicted in Figure 15.

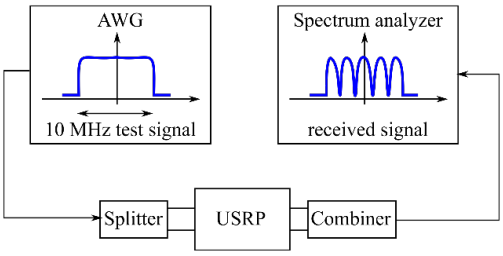


Figure 15: Block diagram of the setup for the verification of the emulated channel transfer function in the frequency domain. An arbitrary-waveform generator (AWG) provides a physical-layer LTE signal with a bandwidth of 10 MHz. The test signal is distributed via a power splitter to the USRP, which sets basic channel parameters. The emulated signal is evaluated with a spectrum analyser.

For N = 2 propagation paths, the channel transfer function emulated by the USRP in the frequency domain results from the Fourier transform of (5):

$$H(f,t) = a_1 e^{j(2\pi[f_{D,1}t - f\tau_1] - \varphi_1)} + a_2 e^{j(2\pi[f_{D,2}t - f\tau_2] - \varphi_2)} \qquad (7)$$

In a first step, a static scenario ($f_D = 0$ Hz) was considered, in order to verify the proper emulation of amplitude weights and path delays. In this case, (7) simplifies to a sum of two complex exponentials, which is entirely determined by the propagation time difference $\Delta\tau = \tau_2 - \tau_1$, the phase difference $\Delta\varphi = \varphi_2 - \varphi_1$, and the amplitude ratio $A = a_1/a_2$. Figure 16 shows the results to the maximum values of the measured transfer function for $\Delta\tau = 500$ ns and $A = \{1, 0.5\}$. Measured and theoretical values agree very well. Corresponding to the inverse of $\Delta\tau$, the separation between two maxima of the transfer function was found to be exactly 2 MHz. In case of equally strong paths (A = 1), deep fading minima emerge. However, the measured data are affected by noise on the one hand, and the limited frequency resolution of 14.5 kHz on the other hand. For A = 0.5, the fading notch amounts to -9.5 dB, and matches the theoretical expectations accurately. After initialisation of the emulator, a random phase shift was observed to occur between the RF frontends of the USRP due to the phase-locked loops, leading to a frequency shift of the transfer function by $\Delta\varphi = 2\pi\Delta f\Delta\tau$. This phase shift was adjusted manually in the emulator settings, since the local oscillators cannot be coupled with the device in use.

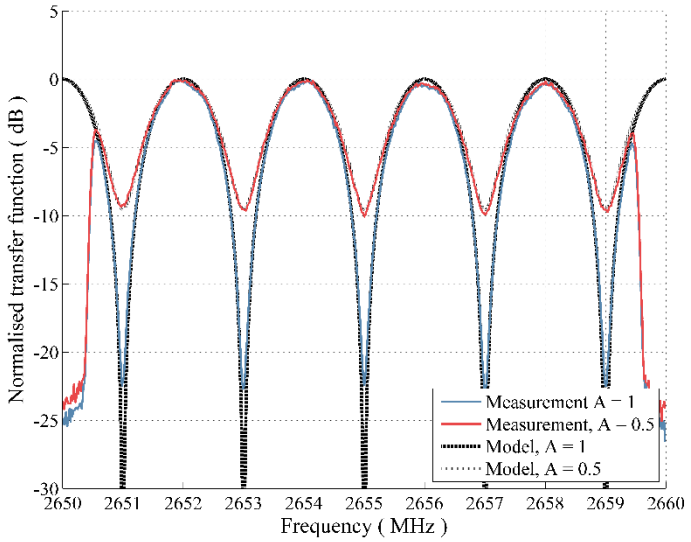


Figure 16: Comparison of emulation results (amplitude ratio of the two paths A = 1: blue curve, A = 0.5: red curve) and analytical data of the static two-path model (black and grey curves). The decay of the transfer function at the edges reflect the band-limited test signal reaching from 2650.5 MHz to 2659.5 MHz.

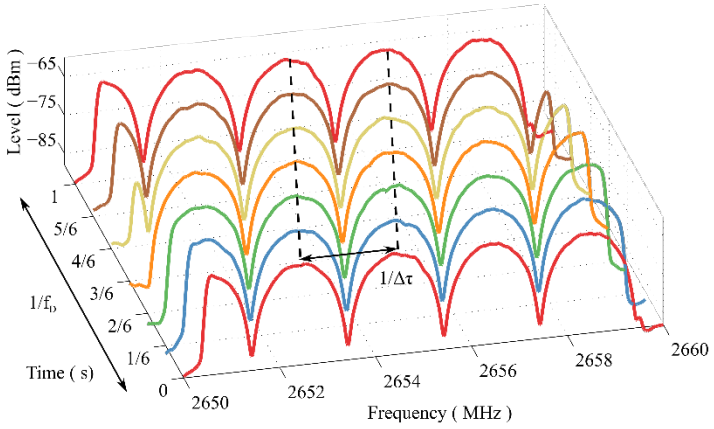


Figure 17: Temporal change of the emulated transfer function, recorded with a spectrum analyser. In case of the two-path model, the shape of the curve remains the same, but is shifted along the frequency axis, until one full Doppler-period is completed.

In a second step, a Doppler shift was applied to one of the two equally strong (A = 1) multipath components, whereas the other path remained static. The transfer function becomes time dependent and has to be recorded in temporal steps, whose resolution is limited by the update rate of the analyser. Figure 17 displays the temporal change of the transfer function for a Doppler shift of 1 Hz. After the time interval $\Delta t = 1/f_D = 1$ s, the initial spectrum was recovered, confirming the correct emulation of the Doppler effect. Higher Doppler shifts up to 500 Hz were verified with similar quality.

4.2. Over-the-air throughput measurements in VISTA

Propagation model of a simplified 2 x 2 MIMO scenario

A simplified 2 x 2 MIMO communication scenario, sketched in Figure 18, provided the basis for the OTA emulation experiment. Beside verification of the emulator on the physical layer in a cable connected setup, its operational capability for end-to-end testing was studied in over-the-air throughput measurements in VISTA. The transmission mode was chosen as open-loop MIMO (TM3),

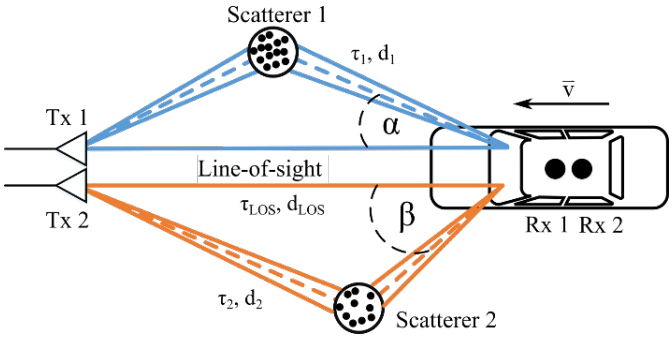


Figure 18: Sketch of the simplified 2 x 2 MIMO scenario with a passenger car driving with velocity v towards a fixed base station. Line-of-sight as well as non-line-of-sight components are considered through different path lengths d_{LOS}, d_1, and d_2 respectively. Blue colour denotes MIMO data stream 1, orange colour denotes data stream 2.

as this is the mode widely used in practical applications. The delays τ_n were chosen from a real drive test in Ilmenau [25]: $\tau_{LOS} = 500$ ns, $\tau_1 = 959$ ns, and $\tau_2 = 1050$ ns. The remaining channel parameters were derived from geometric considerations. With c_0 being the speed of light and $d_n = c_0 \cdot \tau_n$, the relative path weights result from $a_n = a_{LOS} \cdot \tau_{LOS} / \tau_n$. According to the speed of the vehicle and the angle-of-arrival of the impinging waves, the Doppler shift is given by $f_D = v/\lambda \cdot \cos(\text{AOA})$.

Over-the-air measurement setup

The proposed model provides the basis for the evaluation of the achievable data throughput in a driving scenario under realistic channel conditions, emulated by two USRP modules. The angles-of arrival were set to $\alpha = 25°$ and $\beta = 40°$. The standard-compliant Huawei LTE modem mentioned in Sec. 3 was used in combination with the communication tester, to establish a LTE link for throughput measurements; Figure 19 and Figure 20 show the block diagram and the implementation of the setup, respectively. In order to control the two USRP modules with one PC, and to synchronize their internal clocks, additional hardware was required, as explained in relation to Figure 21. The four downlink antennas were arranged on a circle around the test object with a radius of 3.95 m, in a way that the designated angles-of-arrival are emulated. The throughput measurements were conducted in LTE band 7 at a centre frequency of 2655 MHz and with a channel bandwidth of 10 MHz.

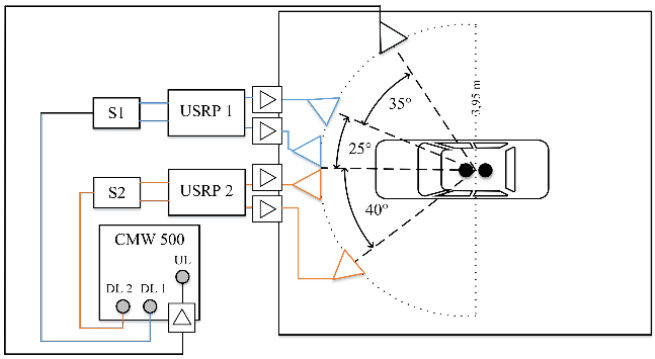


Figure 19: Block diagram of the emulation setup. The communication tester CMW500 provides the LTE protocol stack as well as the downlink RF signals. The downlink signal is distributed to the USRP via power splitters. For the sake of simplicity, the figure is limited to the relevant RF paths. Control signals and conductions for synchronisation are omitted.

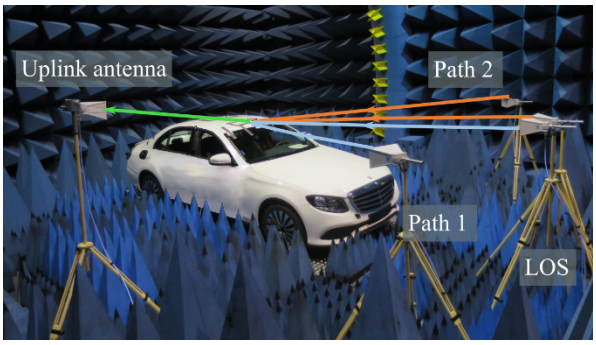


Figure 20: Measurement setup in VISTA for the emulation of the proposed MIMO scenario. The downlink signal paths are coloured blue and orange corresponding to MIMO data streams 1 and 2, respectively (c.f., Figure 18). The uplink path is highlighted in green colour. The downlink antennas are fed by two USRP modules, as indicated in Figure 21.

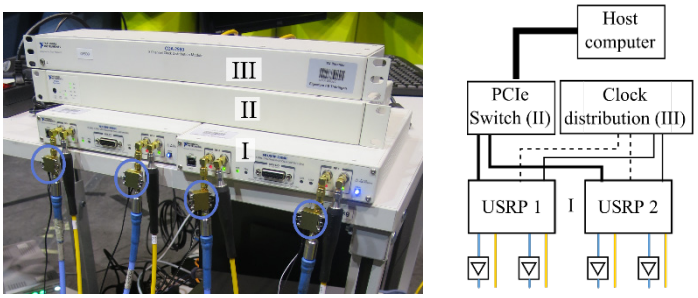


Figure 21: SDR setup for downlink channel emulation. I: The USRP modules perform the digital signal processing. The RF input signal is carried by the yellow cables. After signal processing on the internal FPGA, the signal is amplified (amplifiers indicated by blue circles) and conducted to the illumination antennas via the blue cables. II: PCIe switch for controlling both USRP modules by one host computer. The host computer is connected via a PCIe x8 connection to the switch, whereas the USRP are conducted via x4 connections. III: 10 MHz (right-hand side, solid black line) and 1 PPS (right-hand side, dotted black line) clock distribution in order to synchronize the clock of the both USRP.

Two types of measurement were performed, distinguishing between a reference measurement channel (RMC) mode and the follow-CQI (channel quality indicator) mode. Operating in RMC mode, the modulation and coding scheme is fixed; in this study, a 64 QAM scheme was chosen for the downlink. In contrast, the follow-CQI mode represents a more realistic option as it applies link adaptation, corresponding to the channel quality indicator, which is reported by the user equipment. In both modes of operation, a combination of forward-error-correction (FEC) and automatic retransmission request (ARQ), known as hybrid ARQ (HARQ), was used [26].

Over-the-air Emulation Results

Figure 22 shows the measured throughput values dependent on the received signal level for the static case. The limit of performance of the user equipment, regarding the maximum throughput, is found to be -92 dBm in RMC mode. For input levels below -92 dBm, the throughput dropped drastically. When link adaptation is applied, poor channel conditions are compensated by using a more robust modulation, e.g. 16 QAM or even QPSK instead of 64 QAM, or by reducing the code rate.

In a second step, vehicle motion was emulated, in order to stress the device-under-test when operated at the power limit. Since the LTE downlink uses frequency division access with orthogonal subcarriers (OFDMA), it is expected that Doppler shifts will have a detrimental effect on the performance due to loss of orthogonality (inter-carrier interference). Additionally, synchro-

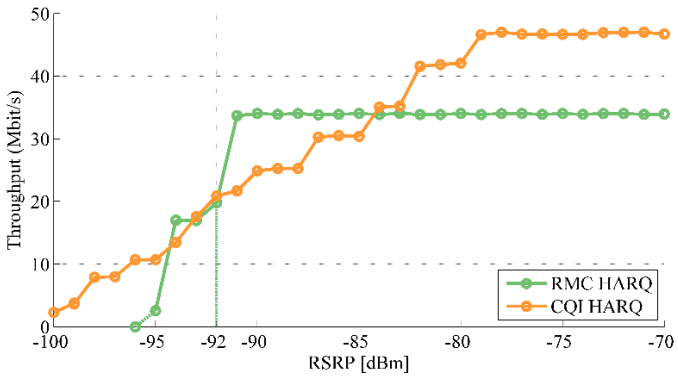


Figure 22: Measured data throughput results versus RSRP. For the reference measurement channel (green curve), the throughput remains constant at 34 Mbit/s until a critical power level of -92 dBm is reached. The data rate drops drastically when the input power level is further reduced. In contrast, when link adaptation is applied (orange curve), the maximum throughput is increased to 47 Mbit/s and the data rate decreases stepwise as channel quality (received power level) decreases.

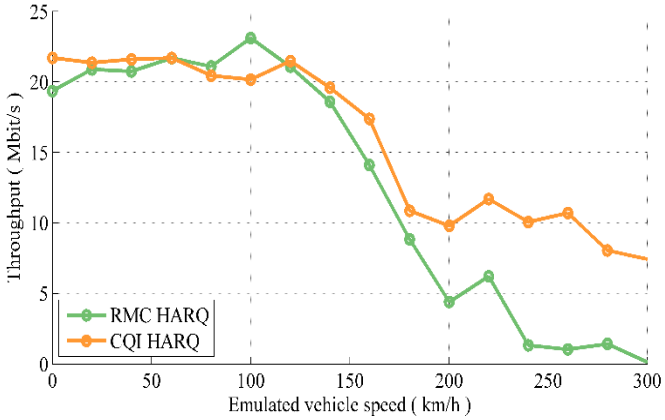


Figure 23: Results for the measured throughput dependent on the emulated vehicle speed. The device was operated at the power limit of -92 dBm. Up to speeds of 150 km/h, the user equipment behaves similarly in both operational modes RMC (green curve) and CQI (orange curve).

nisation failures were also found to affect the throughput, when the device-under-test moved faster than 250 km/h [27]. Figure 23 shows the measured data throughput dependent on the emulated vehicle speed. The input power level remained constant at -92 dBm, while various velocities were tested. Confirmed by the cable-connected measurements, data transmission could be sustained for velocities up to 400 km/h. However, as this is not a realistic benchmark for passenger cars, the diagram is limited to a maximum velocity of 300 km/h, corresponding to a maximum Doppler shift of 740 Hz in the considered LTE band. Shallow variations of the measured throughput occur at emulated speeds below 150 km/h, corresponding to a maximum Doppler shift of 370 Hz, in both RMC and CQI modes. Different behaviour is observed for higher speeds: For the reference measurement channel, the throughput decreases strongly until data transmission fails. With link adaptation, the throughput decreases to a minimum of approximately 7.5 Mbit/s, while the link remains operational.

5. Conclusions

Over-the-air testing is an essential concept for the evaluation of automotive wireless systems, because antennas might not be accessible in installed state due to functional integration. Though knowledge of the radiation patterns of the application antennas remains a key for the evaluation of the radio system, the functionality and functional safety of connected cars require more than mere physical-layer testing. Antenna radiation pattern measurements as well as end-to-end communication testing can be applied in virtual electromagnetic environments like the virtual road simulation and test area (VISTA). In this context, this paper presented a holistic testing concept for automotive wireless systems, entirely based on LTE downlink transmission parameters. It was shown that the gain pattern of automotive antennas can be derived from known reference symbols (RSRP) in the LTE downlink scheme, detected by a commercial LTE communication modem. In contrast to conventional measurement techniques, this approach works without access to the antenna feed, rather utilizing the user equipment as the measurement instrument. The gain pattern of a monopole mounted on the roof of a mid-sized passenger car mock-up was derived and compared to data sets from direct power transmission and conventional nearfield measurements with subsequent nearfield-to-far-field transformation. The three sets of measurement were found fully consistent. The shapes of the pattern showed very good agreement, as indicated by a low variance of the deviations and a high cosine similarity > 96%. The RSRP concept tends to deliver slightly higher gain values compared to the other two approaches. Altogether, the RSRP-based measurement presents a promising effective alternative to conventional techniques, especially when access to the antenna feed is limited due to the installation place, or the integration with the RF frontend.

The end-to-end wireless link performance of the device-under-test was examined in terms of downlink data throughput. While widely used channel models rely on stochastic fading models, this paper proposed a simplified emulation concept based on mapping physical parameters to a set of illumination antennas arranged around the device-under-test. Compared to expensive commercial channel emulators, software-defined radio modules offer a cost-efficient opportunity to emulate basic channel parameters like path delays, Doppler shifts, and the corresponding parameter spreads. The viability of the SDR-based channel emulator was verified by a cable-connected evaluation of the transfer function with a spectrum analyser for the relevant reference scenario of a two-path model. The comparison of the emulated data with analytically calculated reference values revealed very good agreement. By using two USRP modules, the channel emulator was installed in VISTA, and over-the-air throughput measurements were conducted. In addition to low received power levels, rapid channel fluctuations due to Doppler shifts may interrupt the LTE connection since they cause a loss of orthogonality and induce synchronization failures. The measured data throughput dropped for emulated velocities above 150 km/h, when the device operated at the critical power level of -92 dBm for the static case. Beside physical channel conditions and user equipment specifications, the configuration settings of the base station affect key performance indicators like the downlink throughput. In order to simulate the behaviour of the user equipment in a real-world scenario, the application of link adaptation is advisable.

Future work will address the retrieval of phase information of highly integrated antennas from LTE parameters, without accessing the RF feed of the AUT, as well as full three-dimensional recording of the radiation patterns. For end-to-end testing, the proposed setup employing multiple SDR modules will be extended by additional USRP devices, in order to represent a richer multipath environment and thus approximate realistic conditions more closely.

Acknowledgment

The research has been funded by the Federal State of Thuringia, Germany, and the European Social Fund (ESF) under the grant 2015 FGR 0088. The antenna test range VISTA has been funded by the Federal State of Thuringia, the European Regional Development Fund (ERDF), and the German Research Foundation (DFG, 250758117) with financial support from Fraunhofer IIS. The authors would like to thank Mario Lorenz and Dr. Tobias Nowack for their support during the measurements. We further acknowledge funding for the article processing charge by the German Research Foundation (DFG) and the open access publication fund of the Technische Universität Ilmenau.

References

[1] P. Berlt, F. Wollenschläger, C. Bornkessel and M. A. Hein, „Reliable derivation of automotive antenna gain patterns from LTE communication parameters“, 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018
[2] 5G Automotive Association, http://5gaa.org; online, 2018
[3] Car 2 Car Communications Consortium, http://car-2-car.org; online, 2018
[4] Virtual Drive Test Alliance, http://tu-ilmenau.de/vdt-alliance; online, 2018
[5] Z. H. Mir and F. Filali, “LTE and IEEE 802.11p for vehicular networking: a performance evaluation.” EURASIP J. Wireless Comm. and Networking 2014 (2014)
[6] C.A. Balanis, „Antenna Measurements“ In: Antenna Theory Analysis and Design. Wiley, 2005.
[7] J. E. Hansen, Ed., "Spherical Near-field Antenna Measurements", Institution of Engineering and Technology, London, UK, 2008.
[8] Specification USRP-2954, 10 MHz to 6 GHz Tunable RF Transceiver, National Instruments. Available: http://ni.com/pdf/manuals/375725c.pdf; online, 2018
[9] Thuringian Center of Innovation in Mobility, http://mobilitaet-thueringen.de; online, 2018
[10] M. A. Hein et al., "Emulation of virtual radio environments for realistic end-to-end testing for intelligent traffic systems," 2015 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), Heidelberg, 2015, pp. 1-4.
[11] R&S CMW Wideband Radio Communication Tester Specifications, Rohde&Schwarz. Available: https://tinyurl.com/rscmw500; online, 2018.
[12] European Telecommunications Standards Institute, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer – Measurements” *European Telecommunications Standards Institute*, ETSI TS 136 214 V15.3.0 (2018-10), p. 8. Available: http://etsi.org.
[13] Harald T. Friis, "A Note on a Simple Transmission Formula", Proceedings of the I.R.E. and Waves and Electrons, May, 1946, pp 254–256
[14] F. Wollenschläger, S. N. Hasnain, C. Bornkessel and M. A. Hein, "Precision Phase Measurements of Automotive Antennas for Localization in Anechoic Chambers", 12th European Conference on Antennas and Propagation (EuCAP 2018), London, UK, 2018.
[15] A. Stjernman, “Relationship between radiation pattern correlation and scattering matrix of lossless and lossy antennas”, IEEE Elect. Lett., vol. 41, no. 12, pp.678-680, Jun. 2005.
[16] A. Paulus, J. Knapp and T. F. Eibert, "Phaseless Near-Field Far-Field Transformation Utilizing Combinations of Probe Signals," in IEEE Transactions on Antennas and Propagation, vol. 65, no. 10, 2017.
[17] C.H. Schmidt and Y. Rahmat-Samii, "Phaseless spherical near-field antenna measurements: Concept, algorithm and simulation," 2009 IEEE Antennas and Propagation Society International Symposium, Charleston, SC, 2009, pp. 1-4.
[18] D. Smith, M. Leach, M. Elsdon and S. J. Foti, "Indirect Holographic Techniques for Determining Antenna Radiation Characteristics and Imaging Aperture Fields," in IEEE Antennas and Propagation Magazine, vol. 49, no. 1, pp. 54-67, Feb. 2007.
[19] European Telecommunications Standards Institute, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception” *European Telecommunications Standards Institute*, ETSI TS 136 211 V10.0.0 (2011-01), pp. 56ff. Available: http://etsi.org.
[20] C. Schneider et al., "Directional analysis of multipath propagation in vehicle-2-vehicle channels," 2016 10th European Conference on Antennas and Propagation (EuCAP), Davos, 2016, pp. 1-5.
[21] D. Shutin, “Clustering wireless channel impulse responses in angular-delay domain”, 2004 IEEE 5th Workshop on Signal Processing Advances in Wireless Communications, 2004, pp. 253-257
[22] NI, “Real-Time MIMO Channel Emulation on the NI PXIe-5644R,” National Instruments, White Paper, May 2013. [Online]. Available: http://www.ni.com/example/31556/en/
[23] L. Jäger, P. Berlt, C. Bornkessel and M.A. Hein, „Distributed Spataial Channel Emulation for Virtual Drive Testing Based on Multiple Software-Defined Radios“, 2019 13th European Conference on Antennas and Propagation (EuCAP), Krakow, 2019 (accepted)
[24] F.G. Friedlander and M.S. Joshi, “Introduction to the Theory of Distributions”, Cambridge University Press, 1998
[25] J. Gedschold et al., “Tracking based Multipath Clustering in Vehicle-to-Infrastructure Channels”, 2018 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, 2018
[26] E. Dahlmann, S. Parkvall, and J. Sköld, 4G LTE / LTE Advanced for Mobile Broadband. Academic Press, 2011.
[27] D. Micheli, M. Barazzetta, R. Diamanti, P. Obino, R. Lattanzi, L. Bastianelli, V. M.Primiani, and F. Moglie, “Over-the-Air Tests of High-Speed Moving LTE Users in a Reverberation Chamber,” in IEEE Transactions on Vehicular Technology, vol. 67, no. 5, May 2018, pp. 4340–4349

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 311-316 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Talk Show's Business Intelligence on Television by Using Social Media Data in Indonesia

Eris Riso, Abba Suganda Girsang*

Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta,11480, Indonesia

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 09 February, 2019
Online : 28 February, 2019

Keywords:
Business Intelligence
Data Warehouse
Sentiment Analysis
Word2vec
Naïve Bayes
Random Forest

ABSTRACT

Knowing how and types of talk shows discussed in social media is significant to all stakeholders in a talk show's program. There are many messages that can be found in social media that need to be noticed so the messages from the user could reach the viewer. Social media provides promising as well as challenging data to business intelligence. By using data warehouse and business intelligence, the data from social media can be proceeded to monitor and understand adoption and sentiment. In this research, data warehouse and business intelligence will be designed and implemented using data from social medias; Facebook, Twitter, Instagram, and Youtube which are expected to be the source to do monitoring to social media contents that are shared by each stakeholder from the talk shows that are going to be analyzed. As for the sentiment analysis, word2vec, Naïve Bayes and Random Forest will be applied for the qualification methods. As a result, this monitoring is used to give understanding of how information spread and the sentiment which underlies the spreading messages could give critical perception to the management on how their product recognized and valued in society.

1. Introduction

Based on the data of Television Program Quality Index in 2016 Period 5 which is released by Broadcasting Commission of Indonesia (Komisi Penyiaran Indonesia-KPI) that talk show program get scale index 3, 48 out of 4 in according to KPI's standard. This index is higher than news program that gets 3,44, comedy at 3,27, variety show at 3,06, film and tv series at 2,75 and infotainment at 2,71 [1]. Talk show program as a product or commodity is possible to compete with other products. As a pop-culture product, talk show program must be marketable [2]. Through social media [Facebook, Twitter, Instagram, Youtube], communication occurs in real time with a wider coverage area. This program recognized as a place to share ideas and information as well as entertainment in order to give valuable knowledge to the viewer.

Business starts to see technology as an effective mechanism to interact more with their client and customer [3]. The content, time and frequency of conversations in social media about a product between customers are beyond the control of the manager. In that, managers need to learn how to form consumer's discussion in a way which is in line with the mission and target of the organization [4]. The key to success in taking the advantage of consumer involvement is in the understanding of the consumer itself, such as; dissatisfaction, wishes or even their unconscious feeling [5].

At least there are two implications why a company directly involves in social media. First is the need to evaluate social objects and the second one is related to how far a content should be spread so the content could be managed as best as it could, therefore a content management system is needed to identify and manage the content in social media [6]. The Monitoring system is also needed in order to see viewer sentiment, so it could give confidence to the continuity of the talk show program, at least in public perception. With data monitoring, every company that broadcast talk show program could also see the competitiveness level so they could quickly formulate the strategy to face the competition. To date, TV programs are rated by the quantitative survey, thus qualitative survey is also important to be implemented to determine which programs are qualified and worth to be watched by the viewer. This will tell how popular a talk show program in each TV station is and what aspect needs to be improved based on the responses of the TV viewers.

*Abba Suganda Girsang, Email: agirsang@binus.edu

The high amount of information that users continually provides to the Internet requires methods that sort this information and stores it in a way that can be easily accessed and processed [7]. In its original forms, the data contains noise and unstructured that cause difficulty to extract the information in real time. By using data warehouse and business intelligence, the data from social media can be proceeded to monitor and understand adoption and product sentiment. Understanding how the information spread and the sentiment which underlies the messages will give critical perspective to the management about how their product recognized [8]. The application of business intelligence as a tool to monitor could provide reliable information to make effective decision as well as helping in evaluating consumer values and the desire to gain profit [9].

Sentiment analysis is an important tool for the extraction of information about the human emotional state [10]. Since the internet has become an excellent source of consumer review, the area of sentiment analysis has seen a large increase in academic interest [11]. Researches and the use of social media data for sentiment analysis or business intelligence have been done by some people, either for business necessity or else. Machine learning technique can be used to conclude sentiments on social media data that shows the perspective and experience of medicines and cosmetics consumer. The result of this study suggests considering spam comments, comparing the performance of different machine learning sentiment classification, temporary analysis to detect the decreasing tendency or the decline of their sentiment or of a specific product [12]. Decision Support System base on sentiment analysis able to efficiently support companies and enterprises in managing promotional and marketing campaigns on multiple social media channels [13]. Survey analysis shows that people are willing to be involved in the creation of educational policy and their opinion on social media could be directed to the creation of effective educational policy. However, the twitter classification [positive, negative and neutral] has not done in this research [14].

Other research has shown that social media data could be used to form a data warehouse to analyze data in social media such as; likes, comments and sentiment analysis that will be used as the object to design business intelligence to observe the company performance based on the data from their social media. This research will also state that the classification accuracy could be improved by using the algorithm and machine learning like naïve Bayes classification and in the future, Facebook emoticon could also be analyzed to get more comprehensive information [9].

Other research is the analysis of social media content by using some methods of data mining, thus recent information could be collected to support decision making that could give benefit to either company or personal. The objective of this research is to create business intelligence dashboard to observe the performance of news channels that is posted to social media accounts like Facebook and Twitter. This study also suggests wording qualification using Word2vec that could enhance the accuracy of text qualification process [15,16].

In this research a data warehouse and business intelligence software will be designed and implemented by using the data from social media such as; Facebook, Twitter, Instagram, and Youtube which are expected as the sources to do the monitoring towards the contents of those social media which are shared by each stakeholder of talk show programs that are going to be analysed. As for the sentiment analysis, word2vec and Naïve Bayes and Random Forest algorithms will be applied for the qualification methods. So, this research is also expected to give perception of the rating of a talk show program qualitatively.

2. Research Method

Generally, steps in this research are collecting data, sentiment analysis, design of data warehouse, and design of business intelligence. Figure 1 is the proposed model which consists of data collection process, sentiment analysis process, and data warehouse process. The detail of these steps can be seen in the next subsection.

Figure 1: Proposed Model

2.1. Data Collection

This part consists of how the data collected from social media for each official account of the talk shows. Chosen social media for this research are Facebook, Twitter, Instagram, and Youtube. The data is taken by using web scraping method with Node.Js and Python programming language. Collected data will be saved in a file with comma separated value (csv) form. Each web scraping process for each social media account will produce 2 files in CSV format, which is one file containing the admin post and one file containing the user post. Each file with the same type will be combined into one data set file that is used for two purposes. The first is aimed for sentiment analysis process and the second is to be saved in database staging for the necessity of making data warehouse and business intelligence. In this research, will use two years of data, from March 2018 to August 2018.

2.2. Sentiment Analysis

The sentiment analysis will be carried out on this stage. Pre-processing is done in each data set such as tokenization which divides sentences into words or tokens. Then delete mentions, hashtags, and URL. The next is the stopwords process which consists of non-semantic words like article, preposition, conjunction, and pronoun. Pronouns like “he”, “they”, and “I” only have less or no information about sentiments.

After getting the data set from the pre-processing process, it will be continued to the classification process. It is using word2vec to produce word embeddings and Naïve Bayes (NB)

and Random Forest as the machine learning's text classification method. This classification will produce a vector for each word from the data set, then these vectors represent how we use words to describe the relation such as the word "man" to "woman", as "king" to "queen". Representation of words within a sentence can define the characteristics of the words and word embedding of sentences can determine the word characteristics and the context [17].

Talk Show	Facebook	Twitter	Instagram	Youtube
Hitam Putih	-	@HitamPutihT7	officialhitamputihtrans7	TRANS7 OFFICIAL
Ini Talk Show	-	@Ini_Talkshow	initalkshowofficial	Ini Talk Show
Kick Andy	-	@KickAndyShow	kickandyshow	-
Rumah Uya	-	@RumahUya_Trans7	rumahuya_trans7	TRANS7 OFFICIAL
Indonesia Lawyer Club	ILCtv1	@ILC_tvOnenews	indonesialawyersclub	Indonesia Lawyer Club tvOne
Sarah Sechan	-	@SarahSechan_NET	sarahsechan_net	SarahSechanNet
Economic Challenges	ecmetrotv	@EC_MetroTV	-	-
Pagi-pagi	-	@PagiPagi_NET	pagipagi_nettv	Pagi Pagi NET

Figure 2: Social Media Accounts of Talk Show

Word2vec method treats each word equally because the purpose of word2vec is to calculate word embeddings. In word embedding, not all words represent the meaning of a particular sentence. To obtain the feature vector for each review, first, we learn the vector representation of words and then average all vectors of the words in each review. It uses 10-fold cross validation for Naïve Bayes and Out of Bag (OOB) error for Random Forest to fit and validate data training.

2.3. Data Warehouse Process

In designing data warehouse, the Kimbal method will be used where there are 9 stages that must be passed, such as: choosing the process, choosing the grain, identifying and confirming the dimensions, choosing the facts, storing pre-calculation in the fact table, rounding out the dimension tables, choosing the duration of the database, tracking slowly changing dimensions, and deciding the physical design [18].

When the data has been included into the staging database, then the ETL process will be carried out, such as calculating the number of comments, posts, likes, sentiments, and others. Data from the ETL process is then stored in the data warehouse.

Figure 3 is the scheme which is going to be used to acreate a data warehouse and business intelligence which explained that there are 2 fact tables and 6 dimensional tables. The Admin Posts Fact table is a collection of admin data posts history which is done by the admin for each social media account and has foreign key to each dimensional table. Similarly, to Fact User Comment table is the user history comments towards the admin posts for each social media's talk show and has a foreign key to each dimensional table. Meanwhile for each dimensional table is a detailed summary of admin post data and user comments, as well as dimensional time table and dimensional talk show table to complete the needs of the data warehouse manufacture.

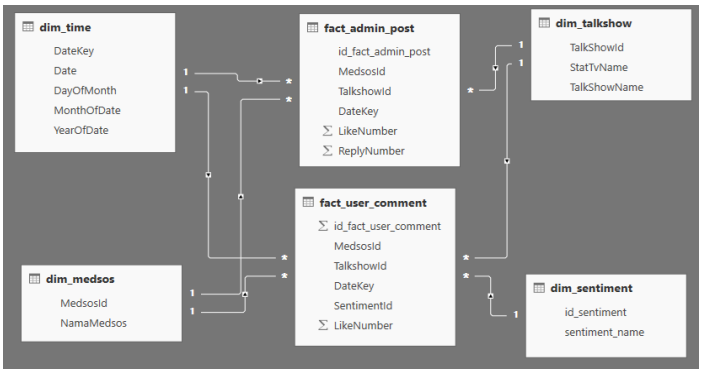


Figure 3: Start Scheme Data Warehouse

2.4. Business Intelligence System

The result of this process is a web-based business intelligence system that can be accessed by the users on the client side. Business intelligence design method that will be used in this research is business intelligence design recommended by Carlo Vercellis. He was written in his book that in developing business intelligence of a company, there are four main phases or stages that must be carried out; analysis, design, planning, implementation, and control [19].

3. Results

Based on the result of the web scrapping that has been done to the data collection process in this research, each talk show's data from each social media's account on period of July 2018 – August 2018 as follows:

Talk Show	Facebook		Twitter		Instagram		Youtube	
	Admin	User	Admin	User	Admin	User	Admin	User
Hitam Putih	1	-	350	635	266	5,182	171	59,947
Ini Talk Show	-	-	610	1,388	215	5,940	68	15,889
Kick Andy	-	-	261	568	114	443	-	-
Rumah Uya	-	-	243	2,035	34	888	277	21,991
Indonesia Lawyer Club	32	2,929	343	4,730	27	859	91	153,057
Sarah Sechan	-	-	11	-	54	683	157	4,264
Economic Challenges	18	26	-	-	-	-	-	-
Pagi-pagi	-	-	2	13	11	163	95	489

Figure 4: Admin Post and User Comment

Figure 4 shows that "Pagi-pagi" talk show has the least admin posts with only have 2 posts in Twitter account and "Pagi-Pagi" talk show has the least user comments with only 13 comments in Twitter account. Meanwhile, for the highest admin post is "Ini Talk Show" with 610 posts in Instagram account and the highest user comment is "Indonesia Lawyer Club" talk show with 153,057 comments in Youtube channel.

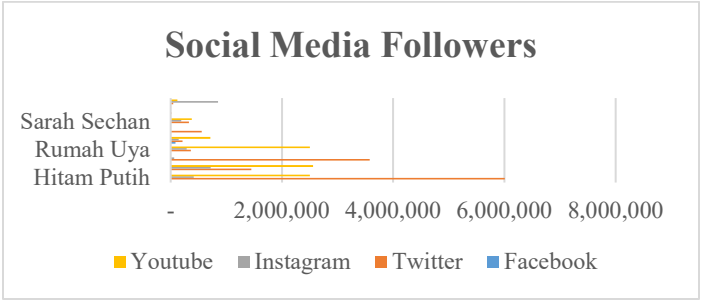


Figure 5: Social Media Followers

On Figure 5 shows the social media followers' data for each social media's account of the talk show. The most talk show's follower is "Hitam Putih" about 6.000.000 followers on social media's twitter, meanwhile, the least talk show's follower is "Economic Challenges" which only has 84 followers on Facebook account

3.1. Sentiment Analysis

In this research, the training corpus for Word2vec contains the comment data for each social's account (Twitter, Facebook, Instagram, and Youtube) from each talk show's programs in Bahasa. The data that has been collected, is the data without label so it needs label to be processed as training data. For the labelling process is used Naïve Bayes algorithm until it gets sentiment positive label, negative, and neutral for each of user's comments. We trained our corpus with 90% of user's comments and each user's comments from each talk show social media account as a testing data. For each of the tests that has been done, 10-fold cross validation is implemented on Naïve Bayes and Out of Bag error on Random Forest.

Talk Show	Facebook			Twitter			Instagram			Youtube		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Hitam Putih	-	-	-	59.24%	59.18%	59.35%	59.01%	59.08%	59.27%	59.14%	59.06%	59.31%
Ini Talk Show	-	-	-	59.28%	59.23%	59.01%	59.08%	59.16%	59.25%	59.08%	59.12%	59.02%
Kick Andy	-	-	-	59.01%	58.94%	59.24%	59.16%	59.14%	59.22%	-	-	-
Rumah Uya	-	-	-	59.12%	59.25%	59.14%	59.26%	59.23%	59.27%	59.17%	59.20%	59.20%
Indonesia Lawyer Club	59.11%	59.07%	59.08%	59.26%	59.11%	59.28%	59.14%	59.13%	59.21%	58.84%	59.14%	59.01%
Sarah Sechan	-	-	-	59.26%	59.00%	59.28%	59.16%	59.08%	59.29%	59.08%	59.25%	59.17%
Economic Challenges	59.18%	58.73%	59.35%	-	-	-	-	-	-	-	-	-
Pagi-pagi	-	-	-	59.04%	59.11%	59.41%	59.17%	59.20%	59.02%	59.39%	58.97%	59.36%

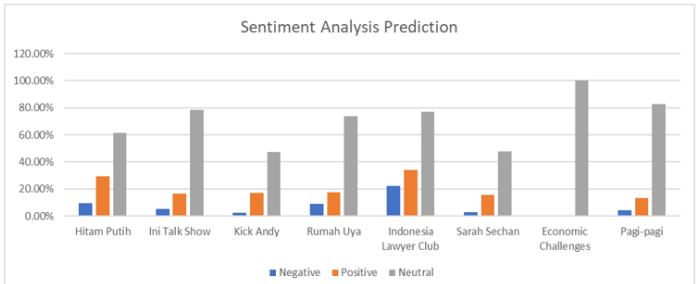
Figure 6: Naïve Bayes Algorithm Test Result using 10-Fold Cross-Validation

Talk Show	Facebook			Twitter			Instagram			Youtube		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Hitam Putih	-	-	-	67.84%	67.95%	68.01%	67.93%	67.84%	67.91%	67.78%	67.94%	67.84%
Ini Talk Show	-	-	-	67.76%	67.95%	68.02%	67.89%	67.83%	67.85%	67.97%	68.01%	67.78%
Kick Andy	-	-	-	67.83%	67.88%	67.91%	67.90%	67.79%	67.74%			
Rumah Uya	-	-	-	67.92%	67.96%	67.76%	67.87%	67.89%	67.78%	67.89%	67.81%	67.64%
Indonesia Lawyer Club	67.73%	67.92%	67.85%	67.72%	67.74%	67.81%	67.90%	67.74%	67.87%	67.78%	67.89%	67.78%
Sarah Sechan	-	-	-	67.91%	68.01%	67.88%	67.85%	67.81%	67.95%	67.72%	68.03%	67.84%
Economic Challenges	67.91%	67.89%	67.68%	-	-	-	-	-	-	-	-	-
Pagi-pagi	-	-	-	67.78%	67.74%	67.76%	67.85%	67.77%	67.89%	67.84%	67.87%	67.83%

Figure 7: Random Forest Algorithm Test Result with Out-of-Bag error

Figure 6 is the summary of Naïve Bayes Algorithm test result as the text classification with cross-validation which the experiments have been conducted 3 times. As the whole test gets the result that the 2nd test gets the highest accuracy score of the overall test which is 59.41%. Figure 7 is the summary of Random Forest Algorithm test result as the text classification with Out of Bag error which the experiments have also been conducted 3 times for each social media's accounts. The highest accuracy score of the overall test is 68,03%. From the 2nd test result of classification algorithm, therefore Random Forest Algorithm test has the highest accuracy score compared to Naïve Bayes.

Figure 8 is the summary graphic of Sentiment Analysis Prediction which can be concluded based on Sentiment Analysis Prediction towards user's comments that have been done accumulatively. Therefore, the highest negative percentage prediction of the talk show is Indonesia Lawyers Club which is 22.16% and for the highest positive percentage prediction is also own by Indonesia Lawyers Club which is 34.17%. Meanwhile, the second position for negative percentage prediction is Hitam Putih which is 9,30% and for positive percentage prediction is 29.47%.



Talk Show	Facebook			Twitter			Instagram			Youtube			Summary		
	Negative	Positive	Neutral	Negativ	Positive	Neutral	Negativ	Positive	Neutral	Negativ	Positive	Neutral	Negative	Positive	Neutral
Hitam Putih	0.00%	0.00%	0.00%	7.10%	31.48%	61.42%	7.57%	27.78%	64.64%	13.22%	29.13%	57.64%	9.30%	29.47%	61.24%
Ini Talk Show	0.00%	0.00%	0.00%	6.66%	18.81%	74.53%	7.37%	17.93%	74.71%	1.18%	12.74%	86.08%	5.07%	16.49%	78.45%
Kick Andy	0.00%	0.00%	0.00%	5.29%	23.10%	71.60%	2.50%	27.50%	70.00%	0.00%	0.00%	0.00%	2.60%	16.87%	47.20%
Rumah Uya	0.00%	0.00%	0.00%	4.62%	14.00%	81.38%	14.76%	26.71%	58.54%	7.42%	11.56%	81.02%	8.93%	17.42%	73.65%
Indonesia Lawyer Club	17.82%	24.46%	57.73%	16.64%	23.89%	59.47%	18.08%	29.10%	52.81%	13.95%	25.06%	60.98%	22.16%	34.17%	77.00%
Sarah Sechan	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.84%	27.66%	64.50%	1.15%	19.84%	79.01%	3.00%	15.83%	47.84%
Economic Challenges	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
Pagi-pagi	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	12.10%	26.75%	61.15%	0.61%	12.47%	86.91%	4.24%	13.07%	82.69%

Figure 8: Summary of Sentiment Analysis Prediction

3.2 Business Intelligence System Development

In order to make easier on monitoring data, Business Intelligence Dashboard is created for admin posts data and for user comments use Microsoft Power BI. Business Intelligence System on admin's side gives facilitation which can be used as monitoring tools, such as the admin post's amount in a certain time scale that is needed. Also, it can be seen how many "like" have been accepted and how many "comments" which have been given by the users towards each admin's posts on each its social medias.

In this monitoring system, it is also displayed sentiment analysis towards each social media such as Facebook, Twitter, Instagram, and YouTube, which are owned by each talk show in order to sell their talk show's product. Besides that, it is also displayed some graphics and which post data that has the most comment and comment displays from each user's comment.

On user's side, this monitoring system displays data which relates with user's comment such as the most active users for each social media's account, sentiment prediction for each user's comment, and which admin's post that is popular for each talk show's users displayed statistically.

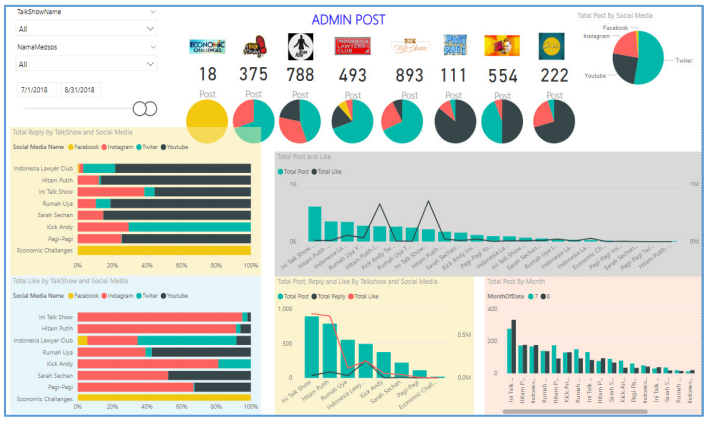


Figure 9: Admin Post Dashboard

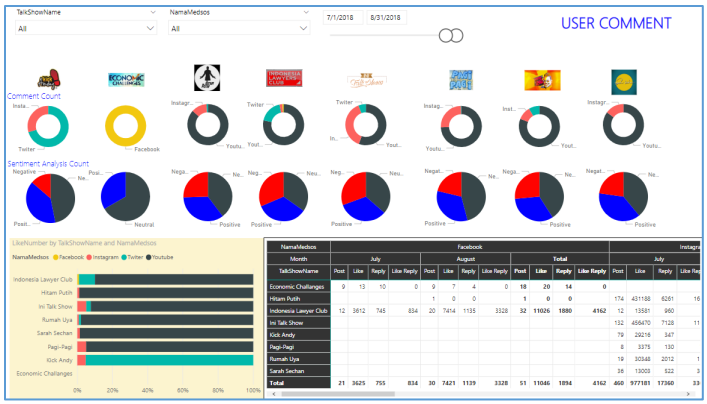


Figure 10: User Comment Dashboard

4. Discussion

In the previous research, sentiment analysis used Word2vec which was more focused on review data with the standard language, so that preprocessing process could be done maximally. In contrast to the sentiment analysis process, it used user's comments data on social media for a talk show which comment did not use standard language. As the result, each of users could give their comments with their own language. Therefore, in the stemming process, there were a lot of words which could not be changed into basic words thus the test scores' results were not maximal.

In business intelligence system that has been made, the data is categorized into 2 categories. They are admin posts' data and user comments' data. For admin's posts category, is made into several pages of data in the form of tables and graphics to display admin's posts, the posts' amount, the popularity of each posts, like, share, total comment, the sentiment's average which is categorized based on talk show's programs and social media's accounts. Meanwhile, for users' comments category is made into several pages which contain tables and graphics to display users' comments data, total comment, like, share, the sentiments' average is also categorized into talk shows' programs and social media's accounts.

#	KPI's Survey 2017	Social Media				
		Followers	Admin Post	User Comment	Like Count	Sentiment Analysis
1	Hitam Putih	Hitam Putih	Ini Talk Show	Indonesia Lawyer Club	Ini Talk Show	Indonesia Lawyer Club
2	Indonesia Lawyer Club	Ini Talk Show	Hitam Putih	Hitam Putih	Hitam Putih	Hitam Putih
3	DR.OZ Indonesia	Kick Andy	Rumah Uya	Rumah Uya	Indonesia Lawyer Club	Rumah Uya
4	Satu Meja	Rumah Uya	Indonesia Lawyer Club	Ini Talk Show	Rumah Uya	Kick Andy
5	Talk to 1 News	Indonesia Lawyer Club	Kick Andy	Sarah Sechan	Kick Andy	Ini Talk Show

Figure 11: Top Five By Number of Viewers and Social Media Activities

Figure 11 shows the five highest data based on the amount of audience which was taken from Broadcasting Commission of Indonesia (Komisi Penyiaran Indonesia-KPI) 2nd period in the year of 2017 (18) and the number of user's activities on social media which based on the result of KPI's survey, the "Hitam Putih" talk show was the most watched talk show. Meanwhile, based on social media data concluded that "Hitam Putih" has the highest score and by this result, we can confirm the KPI's survey result.

5. Conclusion

Generally, Word2vec can be used for sentiment analysis' process in Bahasa, especially in talk shows' programs on television. However, the result of the sentiment analysis did not show positive sentiment or negative sentiment towards its talk show, but it was more likely used as a media to express or comment the opinion freely on social media and it was not meant to give "likes" towards a talk show's program.

Business Intelligence Monitoring System that has been created, it can automatically calculate data from admin's side or user's side which is sourced from the data warehouse. The data such as admin's posts' amount, users' comments, like, share, dislike, who made the comments, who has the most often comments, what are the comments for each of the posts. All of these can be easily displayed in this monitoring system without immediately opening one by one each of the social media's accounts. Therefore, through this monitoring system, the performance can be measured for each of talk show's programs, so that it can be rated from admin's posts and user's comments which talk show's program is popular in social media.

For the next research, particularly for sentiment analysis process, it needs to be developed a stemming method and stop word for nonstandard words especially in Bahasa in order to get basic word precisely from each of user's comments words. Business Intelligence Monitoring System which has been created, it will be better if the data can be processed fast in real time with the use of existing big data technology.

Acknowledgment

The researchers would like to thank for magister programs at bina nusantara university who have supported this research.

References

[1] Komisi Penyiaran Indonesia, "Survei Indeks Kualitas Program Siaran Televisi Periode 5 tahun 2016," 2016.

[2] B. M. Timberg, *Television Talk: A History of The Tv Talk Show*. The University Of Texas Press, 2002.

[3] A. Abdallah, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, "Social media in marketing : A review and analysis of the existing literature," *Telemat. Informatics*, vol. 34, no. 7, pp. 1177–1190, 2017. https://doi.org/10.1016/j.tele.2017.05.008

[4] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," *Business Horizons.*, vol. 52, no. 4, pp. 357–365, 2009. https://doi.org/10.1016/j.bushor.2009.03.002

[5] M. H. Saragih and A. S. Girsang, "Sentiment Analysis of Customer Engagement on Social Media in Transport Online," in 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017. https://doi.org/10.1109/SIET.2017.8304103

[6] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons.*, vol. 54, no. 3, pp. 241–251, 2011. https://doi.org/10.1016/j.bushor.2011.01.005

[7] J. A. Morente-Molinera, G. Kou, C. Pang, F. J. Cabrerizo, and E. Herrera-Viedma, "An automatic procedure to create fuzzy ontologies from users' opinions using sentiment analysis procedures and multi-granular fuzzy linguistic modelling methods," *Information Sciences*, vol. 476, pp. 222–238, 2019. https://doi.org/10.1016/j.ins.2018.10.022

[8] Y. Lu, F. Wang, and R. Maciejewski, "Business Intelligence from Social Media: A Study from the VAST Box Office Challenge," *Comput. Graph. Appl. IEEE*, vol. 34 no 5, pp. 58-69, 2014. https://doi.org/10.1109/MCG.2014.61

[9] M. Yulianto, A. S. Girsang, and R. Y. Rumagit, "Business Intelligence for Social Media Interaction In The Travel Industry In Indonesia," *J. Intell. Stud. Bus.*, vol. 8, no. 2, pp. 72–79, 2018.

[10] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Syst.*

Appl., vol. 118, pp. 272–299, 2019. https://doi.org/10.1016/j.eswa.2018.10.003

[11] Y. Fang, H. Tan, and J. Zhang, "Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness," *IEEE Access*, vol. 6, no. c, pp. 20625–20631, 2018. https://doi.org/10.1109/ACCESS.2018.2820025

[12] H. Isah, P. Trundle, and D. Neagu, "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis," in 2014 14th UK Workshop on Computational Intelligence (UKCI), 2014. https://doi.org/10.1109/UKCI.2014.6930158

[13] P. Ducange, M. Fazzolari, M. Petrocchi, and M. Vecchio, "An effective Decision Support System for social media listening based on cross-source sentiment analysis models," *Eng. Appl. Artif. Intell.*, vol. 78, no. October 2018, pp. 71–85, 2019. https://doi.org/10.1016/j.engappai.2018.10.014

[14] M. S. Omar, A. Njeru, S. Paracha, M. Wannous, and S. Yi, "Mining Tweets for Education Reforms," in 2017 International Conference on Applied System Innovation (ICASI), 2017. https://doi.org/10.1109/ICASI.2017.7988441

[15] P. F. Kurnia and Suharjito, "Business Intelligence Model to Analyze Social Media Information," *Procedia Comput. Sci.*, vol. 135, no. September, pp. 5–14, 2018.https://doi.org/10.1016/j.procs.2018.08.144

[16] P. F. Kurnia, "Perancangan dan implementasi bisnis intelligence pada sistem social media monitoring and analysis (studi kasus di pt. dynamo media network)," Bina Nusantara University, 2017.

[17] M. A.- Amin, M. S. Islam, and S. Das Uzzal, "Sentiment Analysis of Bengali Comments With Word2Vec and Sentiment Information of Words," in 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017. https://doi.org/10.1109/ECACE.2017.7912903

[18] R. Kimball and M. Ross, *The Kimball Group Reader: Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. 2010.

[19] C. Vercellis, *Business Intelligence, Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 314-318 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Emergence of Fun Emotion in Computer Games -An experimental study on fun elements of Hanafuda-

Yuki Takaoka[*, 1], Takashi Kawakami[2], Ryosuke Ooe[2]

[1]*Division of Engineering, Graduate School of Engineering, Hokkaido University of Science, 006-8585, Japan*

[2]*Department of Information and Computer Science, Faculty of Engineering, Hokkaido University of Science, 006-8585, Japan*

ARTICLE INFO

Article history:
Received: 28 August, 2018
Accepted: 29 October 2018
Online: 25 November 2018

Keywords:
Hanafuda
Definition of fun
Questionnaire

ABSTRACT

In recent years, research on game AI has expanded, and now it has become possible to construct even AI of complex games. In accordance with this trend, we constructed the AI of the Hanafuda with a certain degree of complexity. Because of applying the method used in other games to the ball game, we could create a computer player with a certain strength. However, some players feel that strong players are not fun. Therefore, we tried to build a computer player that feels interesting. In the previous experiments, the evaluation for the constructed player was not good. In this research, we conducted a questionnaire survey on players of Hanafuda to raise the evaluation of computer players. The result proved that there are some elements of fun in common among the players.

1. Introduction

Researches on game AI have been actively conducted. Especially in recent years the advancement of computers has made AI for games that has been considered difficult. There are "Mafia (also known as "Werewolf") [1]" and "Poker [2]" in what has been reported. Especially in Mafia, it is proposed to use LSTM (Long Short Term Memory) for analysis of games. Also on poker, it has evolved to defeat Texas Hold'em professional player. In this way, it is possible to create a game AI with a high degree of difficulty and advancement of technology is felt. In response to this trend, we have studied "Hanafuda", which are card games in Japan. The first step was to strengthen the AI of the Hanafuda. Hanafuda is a unique game in Japan, it cannot be said that research is progressing. Therefore, AI which is installed in various Hanafuda game is moderate in ability, never strong. In order to improve AI's ability, we applied UCT (UCB applied to Tree) algorithm to the Hanafuda. As a result, improvement in ability was seen [3].

However, strong AI was not fun AI. Player who battle against AI are sometimes felt boring. Sometimes players do not want to play against AI. Therefore, we aimed to build an AI that made the human player interesting. In order to achieve this goal, definition of "fun" is necessary. Without the definition of "fun", computer players cannot produce entertainment. In our research, we have done so far, we defined the following two points as "fun".

1. increase the variance of get or lose scores
2. adjust the final score of one match to near ±0

These definitions are those that we have devised their own, it was not accurate. A more precise definition is needed when conducting this research. Therefore, in this research, we aim to accurately define the "fun" of the Hanafuda. Specifically, a questionnaire is made to the people who play the Hanafuda, and the result is analyzed.

The structure of this paper is as follows: First, the Hanafuda that this study targeted is described. Secondly, the results of the questionnaire conducted by this research are described. Finally, we analyze the results obtained and describe the policy of future research.

2. Hanafuda

Hanafuda is traditional Japanese card game, and it is the name of the card used in this game. There are 48 cards in this game. Players aim to acquire them according to the rules. Winning or losing is decided with the card taken by the player.

2.1. Game rules

Hanafuda has various rules. The most mainstream among them is "Koi-Koi". Koi-Koi is a game that makes a combination of specific cards. This game is done by two players. Players scramble for cards with each other. Cards can be taken by matching with cards of the same suit. Figure 1. shows Hanafuda cards and correspondence between the card and the month. The rows of cards belong to the same suit. However, the value of the card is different

[*]Corresponding Author: Yuki Takaoka, 7-15-4-1, Maeda, Teine, Sapporo, Hokkaido, Email: 9172001@hus.ac.jp

even within the same suit. Each card belongs to one of “Hikari”, “Tane”, “Tan” and “Kasu”. The value of the card affects when making a winning hand. Figure 2. shows a relationship between cards and classification. Other cards belong to “Kasu”.

In the player’s turn, select a card on player’s hand. If there is card of the same suit in the field, the player can acquire them. At this time, if there are two cards of the same suit in the field, select either one to acquire. If there are three cards, acquire all cards. On the other hand, if there is no card of the same suit in the field, player puts the selected card in field. After the selected card has been processed, excavate the top card of deck, and do in the same way. Up to this point is the turn of the player. At this point, if the winning hand is completed, it is selected whether to continue the game. If player continue, the player aims to create a new winning hand. When player do not continue, player receive points from the opponent player according to the winning hand. If the winning hand is not completed, give a turn to the opponent. The above is a rough rule of Koi-Koi. Based on this, the following section explains the flow of Koi-Koi.

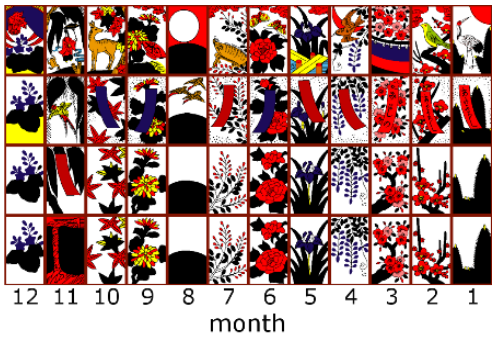


Figure 1. A Hanafuda cards

Figure 2. A relationship between cards and classification

2.2. The flow of Koi-Koi

1. Select the dealer
   First, select the dealer. Selection method is random card draw. Each player draws a card; the player who draws a card close to January is a dealer.
2. Deal the cards
   The dealer deals cards. Eight cards are dealt to each hand and eight cards are dealt to the field.
3. Game start
   A turn is started from the dealer. After that, continue to play until the hand cards runs out or until turn player stops the game on the way. The turn player collects cards according to the rules written in 2.1. If both players run out of cards, give 6 points to the dealer and the next game is started.
4. Game end
   It is one game until cards are dealt and either player gains points. One match is made twelve times in a game. It is the win of the player who has the highest score when one match is over.

2.3. Winning Hans of Koi-Koi

Winning hands of Koi-Koi is as shown in Table 1.

Table 1. Examples of winning hands of Koi-Koi

Name	Combination	Points
Hikari	(an example)	6~
Inoshikachou		6
Shichigosan		6
Omotesugawara		6
Akatan		6
Aotan		6
Tsukimi-Zake		5
Hanami-Zake		5
Tane	omission	1~
Tan	omission	1~
Kasu	omission	1~

Hikari is determined by the combination of acquired “Hikari” cards. Tane is completed by any five “Tane” cards. Tan is completed by any five “Tan” cards. Kasu is completed by any four “Kasu” cards. Each winning hand, one additional point is awarded for every additional card.

3. Previous study

Previously, we conducted research on “fun”. It was to entertain the players by playing the game according to the definitions of interest we defined. In this research, UCT used for game AI was used. In this chapter, we will describe experiments conducted and UCT which is the method used.

3.1. Monte Carlo method and Monte Carlo Tree Search

Normally, the Monte Carlo method is a method of obtaining results by repeating simulation many times. When this technique is used for a game, it progresses the game at random and judges whether the action is good or bad based on winning or losing. Specifically, play randomly according to the rules of the game, and get win or lose. Next, calculate the expected value of the selected action based on the outcome. This flow is performed for all actions that can be selected at a point in the game, and actions with the highest expected value are selected.

Monte Carlo Tree Search (hereinafter called "MCTS") is applied to this method for tree search. Characteristics of MCTS are shown in Figure 3. and Figure 4. The MCTS assigns many simulations to the actions that are considered useful (Figure 3.). And when the number of selections of action exceeds a certain value, the tree is expanded (Figure 4.).

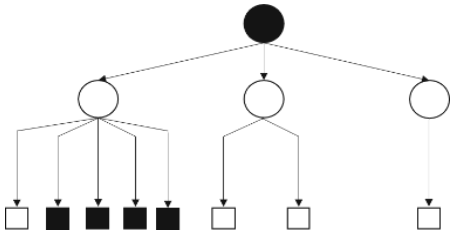


Figure 3. MCTS: many simulations at useful action

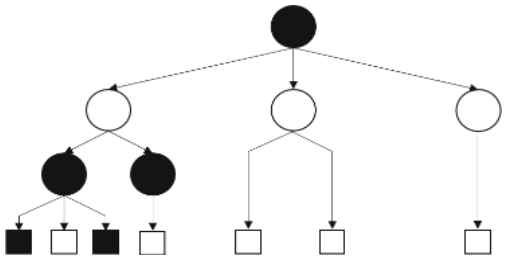


Figure 4. Expansion of tree

The flow of MCTS is as follows.

1. Create a game tree with the current board as the root. In the child node, put behaviors that can be taken at the root node.
2. Proceed the game to the end according to the rules (hereinafter called this act "playout"). At this time, nodes selected a certain number of times are expanded.
3. When winning or losing is confirmed, record it on all the selected nodes.
4. Repeat the specified number of times with 1 to 3 steps as 1 time.

Figure 5. shows these procedures.

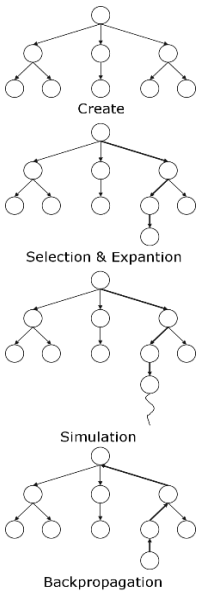


Figure 5. The flow of an MCTS

The advantage of MCTS is, is where it is not necessary design of the evaluation function. By this, it was widely used for games where evaluation of the board surface was difficult. As an example, there is CrazyStone [4] of Go. CrazyStone adopted the MCTS, its ability to win the then Go Tournament at that time.

3.2. UCT

UCT incorporates UCB (Upper Confidence Bound) for tree search. UCB was developed by Auer to solve the Multi-Armed Bandit problem [5]. A commonly used example of the Multi-Armed Bandit problem is a model that gambler plays slot machines. In order to maximize profits, gambler is a matter of thinking which slot machine to play. Gambler uses UCB to solve this problem. The UCB is a value calculated from the play situation of the machine, and by using it, gamblers can make a lot of profit.

UCB is calculated by (1).

$$UCB(i) = \bar{X}_i + c\sqrt{\frac{2\ln n}{n_i}} \tag{1}$$

$\bar{X}_i$

UCT handles each child node as a Multi-Armed Bandit problem and performs a tree search. Update the UCB with simulation results and seek the most valuable behavior.

3.3. Research to produce "fun" by UCT

Among the many definitions of fun, we have defined the definition of "fun" in the previous research as follows. [6].

1. increase the variance of get or lose scores
2. adjust the final score of one match to near ±0

The first definition is to avoid becoming a boring match. It tends to be a boring match if there is little exchange of scores. We decided that match will be interesting if player makes active exchange of score, and adopted this definition. The second definition is to improve the impression of the match. Human players tend to worsen the impression of match if they lose a lot to computer players. Conversely, even if a human player wins too much, it is not good and it is necessary to balance. Therefore, we judged that we should adjust the score to around ± 0 without major win or loss. We adopted this definition for the time being because we got a response that suggested this definition to several players and that it is reasonable.

In order to produce the "fun" that we defined, we attempted to change the usage of UCT and realize it. Specifically, change the decision method when selecting a card in hand as follows.

$$Select\ hand = \begin{cases} \min_i |point + UCB(i)| & point > 0 \\ \max_i UCB(i), & otherwise \end{cases} \tag{2}$$

In (2), *point* is the score obtained by the computer player, and *i* is the number of the hand cards. Equation (2) selects a card that loses score if the score of the computer player is plus. For example, the score of the computer player is +7 points. At this time, the computer player selects a card which can be brought closest to ±0. In this way we tried to produce "fun".

However, the experiment result was not good. The comments of the human players who fought against the computer player are shown in Table 2.

Table 2. Human player's comment

Name	Good point	Bad point
A	• It is not too strong, but it will not let me win easily. • Not a one-sided match, I can enjoy a match to a certain extent.	• Computer players sometimes do strange selection.
B	• I could see it trying to make a game.	• The strength was not constant. I felt it was often too weak.

Analyzing these comments, the following can be said.

- It is not so strong that human player cannot win.
- Computer player is trying to entertain them.
- It feels too weak for some people.
- There are scenes in which selection is considered strange.

Accordingly, "fun" defined by us is insufficient, and more precise definition is necessary. In this research, as a first step for accurate definition, it is to investigate the "fun" that the players of the Hanafuda think.

3.4. Other research

Ikeda et al. 's research is an example of research that produces fun. Ikeda et al. researched "computer Go which entertain human players". For that purpose, Ikeda et al. interviewed experts "What are the elements entertaining Go?" [7]. From the results, researches on elements such as "What is humanness" are under way. Finally, Ikeda et al. said that discovery about "AI like human" was obtained, and they are thinking whether I can create "interesting AI" using this. We decided to conduct research along this trend. As in the case of Ikeda et al.'s research, this research is conducted as a first step to narrow down the elements of fun.

4. Questionnaire survey

In this chapter, the questionnaire survey conducted and the results are described.

4.1. Survey method

We sent a question by e-mail to the person who consented to the investigation and carried out a survey by getting it sent back. The implementation period was from 16th March to 2nd April 2018. There were 76 people who agreed to the survey, 61 of whom returned the mail within the period, and the response rate was 80.3%. Therefore, the total number of respondents is 61 people. In addition, this questionnaire survey is targeted to players who regularly play Hanafuda, and it is not a questionnaire after collecting participants and making a match.

4.2. Survey content

We made the following items as questionnaire contents:

1. Attributes of survey target
   - A. Age and gender
   - B. Number of years that they've been playing Hanafuda
   - C. Frequency of playing Hanafuda
2. Questions about fun
   - A. Elements making the Hanafuda interesting
   - B. Elements making the Hanafuda uninteresting
   - C. Recommendations for Hanafuda

Question 2 is a form of free description.

5. Questionnaire results

5.1. Attributes of survey target

A. Age and gender

Regarding gender, all respondents were male. The age was wide, ranging from teens to 50's, with an average age of 31.9 years.

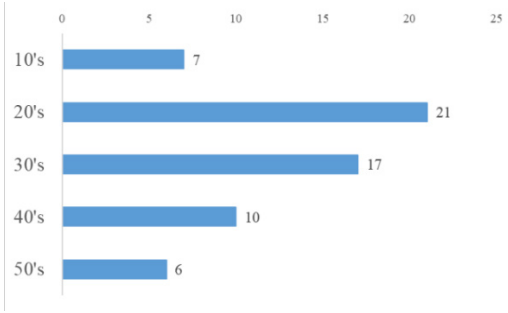


Figure 6. Age of respondent

B. Number of years that they've been playing Hanafuda

The minimum was one year and a half, and the maximum was 15 years.

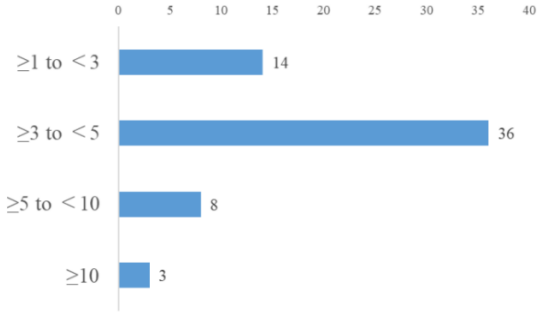


Figure 7. Play number of years

C. Frequency of playing Hanafuda

The most frequent answers were about once a week. The next most frequent answer was about once a month, under it 3 to 4 times a week. In addition, there was an answer that it does not go regularly but does it at a specific time. Incidentally, this answer

includes not only using actual cards, but also those play on electronic devices.

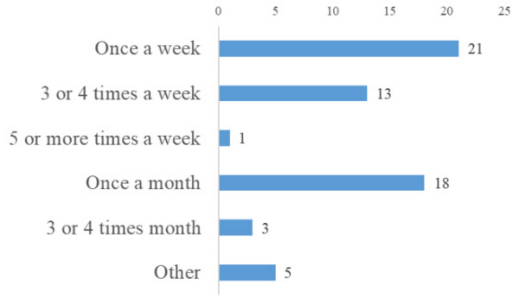
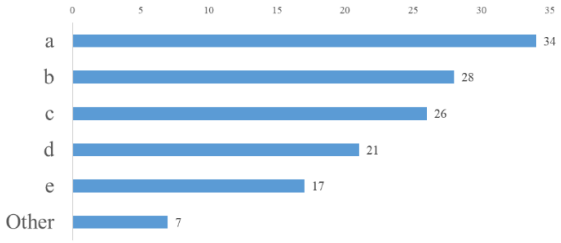
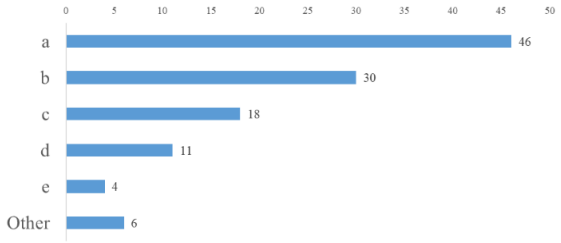


Figure 8. Frequency of playing Hanafuda



a: Completely victory to opponent
b: Got a big score
c: Won by reversing
d: Dodged opponent's attack
e: Interfered the opponent

Figure 9. Elements making the Hanafuda interesting



a: The game ends at an early stage
b: Could not win the game even once
c: The score gained by the opponent is low
d: Drew cards are bad
e: Dealt cards are bad

Figure 10. Elements making the Hanafuda uninteresting

5.2. Questions about fun

A. Elements making the Hanafuda interesting

We asked the fun elements of the Hanafuda. The results are shown in Figure 6.

Elements with many responses are as follows:

- Completely victory to opponent
- Got a big score
- Won by reversing

These items were the top 3.

B. Elements making the Hanafuda uninteresting

Contrary to 1., we asked the suffering elements of the Hanafuda. The results are shown in Figure 7.

C. Recommendations for Hanafuda

We asked about the recommendations for the Hanafuda. The most frequent content is about "Tskumi-Zake and Hanami-Zake". Tsukimi-Zake and Hanami-Zake can get 5 points with two cards. Therefore, the degree of difficulty is low, and the player can easily make a Winning-Hand. It seems that there are many players who view this point as a problem. The second frequent answer was about the rules of Hanafuda. There are various rules in the Hanafuda, but what is mainstream now is "Koi-Koi". There seems to be some players who wish that not only this but also more diverse rules become common. Other, many recommendations on rules and dissemination were received.

6. Future research policy

Future research will conduct a detailed questionnaire survey based on the obtained answers. In the present situation, only the element of the "fun" of the Hanafuda that each player thinks were obtained, and it is necessary to seek elements that satisfy many players. As a specific procedure, we conduct a questionnaire survey of 2 choices for each element obtained in this research. For example, "Yes / No" answers to the question "Do you feel that "Completely victory to opponent" is interesting?". Based on the questionnaire survey, we aim to create computer players that many players feel interesting. It is also necessary to ask many players questions that can refine questionnaire questions and analyze the essence of interest. After conducting detailed questionnaires, detailed analysis is also required.

7. Conclusion

In this research, we investigated interesting elements of the game with the goal of creating a computer player that makes human players fun. Among various kinds of games, this research targeted the Hanafuda that are games in Japan. Survey was conducted by sending questions to players who play Hanafuda. By the reply sent back, we got the fun that each player thinks. In future, it is necessary to further investigate and seek more definite definition of fun.

References

[1] M. Kondoh, et al., "Development of Agent Predicting Werewolf with Deep Learning" in International Symposium on Distributed Computing and Artificial Intelligence, Toledo Spain, 2018. https://doi.org/10.1007/978-3-319-94649-8_3
[2] M. Moravcík, et al., "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker", Science, 356(6337), 508-513, 2017. https://doi.org/10.1126/science.aam6960
[3] Y. Takaoka, et al., "A study on strategy acquisition on imperfect information game by UCT search" in the 2017 IEEE/SICE International Symposium on System Integration, Taipei, Taiwan, 2017. https://doi.org/10.1109/SII.2017.8279334
[4] R. Coulom, "Monte-Carlo Tree Search in Crazy Stone" in Game Prog. Workshop, Tokyo, Japan, 2007.
[5] P. Auer, et al., "Finite-time Analysis of the Multiarmed Bandit Problem", Machine Learning, 47(2-3), 235-256, 2002. https://doi.org/10.1023/A:1013689704352
[6] Y. Takaoka, et al., "A Fundamental Study of a Computer Player Giving Fun to the Opponent", Computer Science & Communications, 6(1), 32-41, 2018. https://doi.org/10.4236/jcc.2018.61004
[7] M. Yamanaka et al., "Bad Move Explanation for Teaching Games with a Go Program", IPSJ SIG Technical Report, 2016-GI-35(5), 1-8, 2016. (Japanese)

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 319-332 (2019)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

An Innovative EPC Architecture based on Not Only Stack for beyond 5G Mobile Networks

Binwei Wu[1], Lu Ge[*,2], Jie Zeng[2], Xiangyun Zheng[1], Youxi Tang[1], Xin Su[2]

[1] *National Key Laboratory of Science and Technology on Communication, University of Electronic Science and Technology of China, 610054, China*

[2] *Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, 100062, China*

ARTICLEINFO

Article history:
Received: 14 December, 2018
Accepted: 02 April, 2019
Online: 18 April, 2019

Keywords:
EPC
Network architecture
Not only Stack framework
Virtualization
Beyond 5G (B5G) networks

ABSTRACT

The traditional long-term evolved (LTE) network architecture faces major challenges due to the rapid development of mobile intelligent applications. Evolved packet core (EPC) network, as the core network of the LTE system, faces the same problem as well. In this paper, we propose an innovative EPC architecture, i.e., Not only stack (NOS)-EPC, in which we redesign the control paradigm based on the NOS framework. NOS-EPC contains the user plane (U-plane), the control plane (C-plane), the management plane (M-plane), and the global network view (GNV). By using NOS framework, distributed control paradigm is turned into a centralized manner which simplifies the signaling flows, reducing the procedure latency and overheads. The operation of NOS-EPC is expounded by four main EMM procedures, including initial attachment, service request, handover, and detachment. Through comparison to LTE/EPC, and software-defined networking(SDN)-EPC, we elaborate the advantages of NOS-EPC on operating expense (OPEX), scalability, flexibility, signaling overheads, and user traffic transmission capability. The simulation results confirm that the proposed NOS-EPC takes advantages on the procedure latency and signalling overheads compared to the other EPC solutions, e.g., LTE/EPC and SDN-EPC.

1 Introduction

Current cellular network (i.e., LTE/EPC) is built as a hardware-based, closed, and inflexible architecture. It faces major challenges due to the rapid development of mobile intelligent applications. It is anticipated that the new generation cellular network, e.g., beyond 5G (B5G) network [1], achieves over 100 and 1000 times enhancement on the aspects of capacity and transmission rate. Advanced wireless techniques are developed to relieve the burdens on the air interface of the cellular network, e.g., massive multiple-input-multiple-output (MIMO), non-orthogonal multiple access (NOMA). However, if we simply rely on these air interface-related techniques, severe challenges in B5G cannot be completely solved since various scenarios pose different requirements. For example, the ultra-reliable low-latency communication (uRLLC) scenarios mainly appeal to the reliability and latency of transmission, while the machine-to-machine (M2M) considers more about the massive volumes of connections. Thus, advanced network means and flexible network elements (NEs) design need to be applied to improve the system performance on the network level.

Considerable works have been done to improve the system performance. One of the most effective ways is the virtualization [2]. Several mainstream virtualization solutions have been proposed, such as network function virtualization (NFV), software-defined networking (SDN), cloud computing [3]. Among these solutions, NFV and SDN are promising approaches [4]. The NFV technique decouples the NEs from the underlay physical infrastructures. NE-special hardware is replaced with the commodity servers, which largely decreases the procure cost. Moreover, NEs could migrate among multiple commodity servers as software

[*]Corresponding Author: Lu Ge, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, 100062, China, l_ge@mail.tsinghua.edu.cn

https://dx.doi.org/10.25046/aj040241

instances. Thus, operators could remotely schedule the resources according to dynamic workloads, decreasing the operational cost. SDN is another promising solution which separates the control plane and data plane. Most SDN-based architectures consist of a distributed data plane and a centralized control plane. The control functions in LTE/EPC are packaged as SDN applications. The SDN controller schedules the network resources in a centralized manner, which improves the programmability of the system.

However, SDN and NFV are originally intended for wired transmission networks. The direct integration of SDN and NFV into cellular networks lead to degeneration of performance improvement. There are several fundamental differences between the cellular network and conventional wired networks. First, in the cellular networks, NEs have more complicated functions compared to the wired network. The control functions include mobility management, traffic engineering, QoS guarantee, etc. The user functions include packet forwarding, integrity checking, etc. On the one hand, existing SDN protocols (e.g., OpenFlow) cannot deliver so many services. What's more, a centralized SDN controller usually results in severe network congestion since it could not support so many services at the same time. Second, the protocol stacks in the cellular networks are tightly coupled. In the wired network, the routing table loosely decouples the routing function and data forwarding function. However, in the cellular network, it is difficult to decouple the control plane and user plane without changing the current protocol stacks. Additionally, each protocol entity in the cellular network maintains complex states (i.e. registered, idle, and connected) that involves a mass of signaling interactions, while the wired network works in a stateless manner.

Therefore, in order to improve the performance of virtualization techniques in the cellular network, it is necessary to redesign the control paradigms in LTE/EPC. In our previous work, we proposed an innovative EPC architecture based on the NOS, i.e., NOS-EPC[5]. The NOS-EPC simplifies the signaling flows in LTE/EPC, reducing procedure latency and signaling overheads. In this paper, we conduct the research in a more comprehensive manner. The main contributions are summarized as follows:

- We propose an innovative EPC architecture based on NOS framework. We detail its evolution from a conventional LTE/EPC. The main components of NOS-EPC are elaborated.
- We use four main EMM procedures to illustrate how NOS-EPC works. The EMM procedures include initial attachment, service request, handover, and detachment. Additionally, we elaborate on how to incorporate new services in NOS-EPC to address the flexibility of NOS-EPC.
- We show advantages of NOS-EPC through comparisons among LTE/EPC, SDN-EPC, and NOS-EPC. We emphasize the features of NOS-EPC on operating expense (OPEX), scalability, flexibility, signaling overheads and user traffic transmission capability.
- We evaluate the performance of NOS-EPC through NS3 simulations. The simulations show that the proposed NOS-EPC could improve system performance significantly, compared to traditional LTE/EPC and SDN-EPC.

The rest of paper is organized as follows. In Section II, we review the current works on the architecture of the cellular network. In Sections III, we describe the structure of our proposed NOS-EPC. Section IV details how the NOS-EPC works, using four main EMM procedures. Then, in Section V, a comprehensive comparison between LTE/EPC, SDN-EPC, and the proposed NOS-EPC is given. We conduct NS3 simulations in Section VI, which evaluates the performance of the proposed NOS-EPC. We conclude the paper in Section VII.

2 Background and related works

The architecture of the cellular network is a hot topic in the past few decades. The current LTE/EPC could not satisfy the strict requirements of 5G and B5G [3]. Considerable efforts have been made to investigate the disadvantages of the current LTE/EPC system.

First, in the LTE system, NEs are built on the dedicated hardware, which raises CAPEX and OPEX. Second, NEs in LTE are highly integrated. In order to deliver a particular service, NEs needs to cooperate with each other which motivates unnecessary singling overheads. Moreover, highly-integrated NE design makes it hard to incorporate emerging applications. Third, diverse protocols are designed for different NEs. Heterogeneous protocols result in extra cost during traffic/signaling transmission.

In order to overcome above-mentioned challenges, some researches focus on the optimization of network architecture. Most works use virtualization techniques (e.g., NFV, SDN) to enhance network programmability. In these architectures, NEs are instantiated as packages on the SDN controllers or the software instances hosted on the VMs[6, 7, 8, 9]. The authors of [6] present an SDN-based architecture, named MobileFlow, to validate the feasibility of using SDN in the cellular network. [7] proposes a cloud-based architecture, i.e., CONCERT, which coordinates physical resources as virtual resources. The authors of [8] present an SDN-based 5G cellular system, i.e., SoftAir, in which a unified U/C-plane interface is designed based on OpenFlow. Another high-level cloudlet-based architecture is proposed in [9]. The authors address the seamlessness and low end-to-end latency between UEs and its Avatar.

EPC, as a part of the cellular system, has also been extensively studied [10, 11, 12, 13]. The authors of [10] propose a programmable mobile core network with an OpenFlow-enable data plane. The control function is

implemented with an OpenFlow Controller that cooperates with MME. [11] presents a number of different options which are associated with the implementation of EPC over a cloud infrastructure and providing it "as a Service". In [12], the authors detail an OpenFlow-based protocol to enhance network flexibility and programmability. A much more detailed analysis is given in [13], in which authors evaluate the network performance, considering five commonly used procedures.

Some other works conduct the optimization from the aspects of a particular service. [14] investigates an SDN/NFV-based platform in order to improve the performance of video broadcasting service. The improvement owes to the orchestrator in the core network, which is in charge of flow operations like replication and transcoding. [15] designs a 5G user plane function component (UPF) based on SDN to support network slicing, reducing the latency and improving the throughput. [16] focuses on PDN data transmission, in which the authors analyze the cost incurred by different virtualization schemes. In [17], the authors propose an SDN-enabled EPC to carry out inter-domains handover, which uses OpenFlow (OF)-enabled switches as the main forwarding nodes.

It is noticed that the above-mentioned researches do not change the control patterns and interfaces in traditional LTE/EPC. The problems brought by the ossified protocol stack in LTE/EPC still exist. Thus, some researchers focus on the redesign of traditional paradigms in LTE. In [18], C-plane of RAN was reconstructed via data-signaling separation in the air interface. The control coverage and the data coverage are separately controlled according to the on-demand network. A virtualized RAN based on NOS is proposed in [19], which is aimed to optimize flexibility and sustainability. Meanwhile, signaling overheads and service response time are reduced.

The redesign of control paradigms in RAN motivates us to conduct similar methods to LTE/EPC. [20] only focuses on the reconstruction of S/P-GW in EPC. In this paper, we consider the whole region of the EPC/LTE network. We reconstruct the control paradigm based on NOS framework to decrease the signaling overheads and latency.

3 An innovative NOS-based EPC architecture for the mobile network

Unlike the mainstream virtualization techniques like SDN and NFV, the NOS focuses on the redesign of the paradigm of the cellular network [19]. It offers a mechanism that can help to build a virtualized cellular network architecture. In this architecture, network resources can be dynamically allocated and configured to improve the efficiency of utilization. Signaling interactions can be optimized to shorten the procedure duration and decrease the network overheads.

3.1 An Innovative EPC based on NOS framework

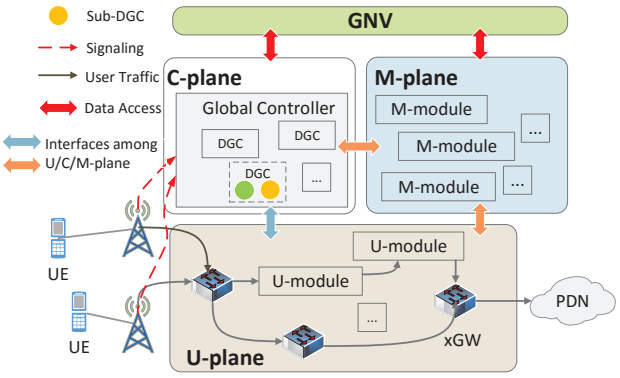


Figure 1: The architecture of the NOS-EPC

The essential methodology of NOS framework is breaking the boundary of protocol entities into different functions, and reconstructing the paradigm of the cellular network in the level of functions. The NOS framework decomposes the functions of NEs into the user plane (U-plane), the control plane (C-plane) and the management plane (M-plane). Data sources which are originally distributed among different NEs are now gathered into a logically centralized GNV. We redesign the network architecture by using NOS framework. The proposed NOS-based architecture is shown in Figure 1. It consists of U-plane, C-plane, M-plane, and GNV. The details of C/U/M plane and GNV are illustrated as followings.

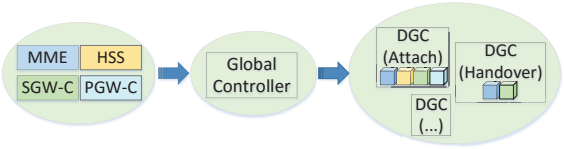


Figure 2: The generations of the DGCs

The U-plane is built in a distributed manner. In the U-plane, the user functions (i.e. GTP encapsulations, rules installations) of the service gateway (SGW) and PGW are merged into the xGW. U-plane modules (U-modules) for emerging services are introduced. Each U-modules is a processing element. The user traffic is routed among these U-modules under the control of C/M plane.

In the C-plane, the traditional control functions are logically centralized into an entity named as GC. The GC is decomposed into multiple modules, denoted as distributed-GCs (DGCs). The splitting of DGCs is conducted according to LTE EMM procedures. Figure 2 demonstrates the process of the generations of these DGCs. The circles with different colors represent the functions in different traditional NEs. SGW-C and PGW-C stand for the collections of control functions in SGW and PGW, respectively. Moreover, the DGCs could be further split up into sub-DGCs in order to decrease its complexity. In order to cover a wide area,

copies of DGCs and sub-DGCs are scattered in different clouds/cloudlets. The states and correlated information of DGCs/sub-DGCs are collected into GNV.

The M-plane is designed for the management and administration of multitudinous DGCs/sub-DGCs and U-modules. Its responsibilities could be categorized into resource management, orchestration, scheduling, interpretation and diagnostic, each of which is implemented as M-plane modules (M-modules). The resource management module is in charge of resource allocation for the NOS-EPC components, such as DGCs/sub-DGC and U-models. These modules (DGCs/sub-DGC and U-models) are instantiated as software applications and hosted in VMs. Thus, the resource management module is used to manage the resources for these components, i.e., create/migrate/delete VMs, etc. DGCs/sub-DGCs and U-modules are orchestrated by the orchestration modules to deliver services. The interpreter modules provide unified Application Programming Interfaces (APIs) for the network operators, reducing the cost of the maintenance and upgrades. The diagnostic modules could automatically detect the system performance issues, providing a decision basis for the scheduling modules or network operators.

GNV contains all the information of users and system. It is logically centralized and could be deployed in a distributed manner with the purpose of reducing the access loads. GNV fully decouples the U/C planes and make the modules in C/U/M planes into stateless. Supported by a logically centralized data source, control functions, and management are easy to be realized. Moreover, global optimization is more convenient to be carried out in NOS-EPC since global network information (view) is provided. From the perspective of implementation, GNV is essentially a distributed database which stores the system/user information. Modules in different plane use the data in GNV as a basis to carry out their jobs. Thus, the performance of NOS-EPC highly depends on the reading/writing speed of GNV. Fortunately, there are plenty of mature techniques in the field of distributed databases. Some of them have already been brought into the commercial market for decades[21, 22].

Unlike the traditional EPC, a unified communication protocol is employed for the interactions among the modules in C/U/M planes. Transaction operations are introduced in the C/M planes to ensure integrity. The multicast and concurrent multipath transfer (CMT) are used.

3.2 Implementation and virtualization of NOS-EPC

The deployments of the physical infrastructures and logical elements in Section 3.1 are discussed in this section. The infrastructures used in NOS-EPC could provide a platform, capable of hosting 5G functions. The infrastructures are classified as commodity servers and switches/routers. All the logical elements (DGCs/sub-DGCs, U/M-modules, and GNV) are hosted within virtual machines (VMs) instantiated on commodity servers. In order to cover a huge area, duplications of NEs (i.e., GNV, sub-DGCs) are installed on different sites. Since the network state, user information, and session information are stored in GNV, it is convenient for the M-models to balance the workloads among multiple duplications.

Ephemeral states of the VMs (i.e. resource utilization, workloads) are synced into GNV. The sync is performed periodically or triggered by events, controlled by M-modules. The placements of logical modules and the associations between modules (DGC/sub-DGC, U/M-modules) and users are dynamically optimized. Switches/routers in the NOS-EPC are extended, similar to OpenvSwitches. Flows including control flows and user flows are identified by FLOW_IDs. The Switches/routers route the user flows according to the flow tables. Flow tables could be dynamically configured by DGCs/subDGCs or M-modules through open APIs.

4 The operations of NOS-EPC

In this section, we illustrate the operations of the NOS-EPC. An unambiguous view of the C/M plane is demonstrated through several user cases. We mainly focus on the initial attachment, detachment, service request, and X2 handover procedure in NOS-EPC. Moreover, to emphasize the flexibility of the proposed NOS-EPC, an example of how to incorporate an emerging application is also demonstrated.

4.1 Initial attachment

4.1.1 Initial attachment procedure in LTE

Initial attachment is the first control procedure for users to access into the LTE network. It is triggered by the attach request sent by the user. The procedure can be divided into five sub-phases according to their different specific purposes which shows as below:

Phase 1: *Mobile equipment identity acquisition (MEIA)*. In this phase, the user equipment submits its equipment identity and the other equipment-related information such as supported secure algorithms and integrity protection algorithms to the network.

Phase 2: *Authorization*. In this phase, users and networks identify each other and verify the integrity protection algorithms through the negotiation.

Phase 3: *NAS secure setup*. In this phase, users and networks confirm the security mechanism with each other to ensure the safety and reliability of exchanging of NAS messages.

Phase 4: *Location update*. In this phase, the users are registered in the network and MME gets the user information related to the service type.

Phase 5: *Default EPS bearer setup*. In this phase, the network assigns corresponding resources to establish communication bearers for the users. These bearers allow users to use some basic services.

The details of the messages exchanging during the procedure vary depending on the circumstances. For simplicity, we assume that neither the user nor network has the context about the last attachment. The signaling interactions during the initial attachment in LTE/EPC are shown in Figure 3 [23].

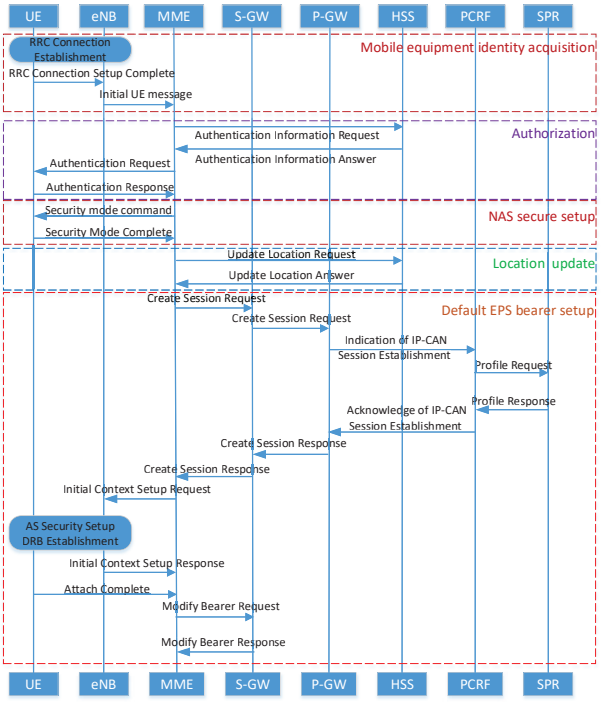


Figure 3: The initial attachment procedure in the LTE/EPC

4.1.2 Initial attachment in NOS-EPC

There are 4 types of sub-DGCs (named as InitialCtrl, AuthCtrl, SecuCtrl, and BearerSetup) involved during the initial attachment procedure in NOS-EPC. The InitialCtrl, AuthCtrl, and SecuCtrl cover the former three phases (i.e., MEIA, Authorization, and NAS secure setup). The BearerSetup covers the latter two phases. The resources (i.e. VMs' CPU and memory) and the dependencies for the DGCs/sub-DGCs are managed by the M-models, which ensures that the user signaling is processed with a right order.

The M-modules allocate the resources (i.e. VMs CPU and memory) for the above sub-DGCs. Dependencies among DGCs/sub-DGCs are determined by the orchestration module, which ensures that user signaling is processed with a right order. In traditional LTE/EPC, the functions related to mobile equipment identity acquisition and NAS security setup phases are centralized in the MME. Therefore, InitialCtrl and SecuCtrl simply play the roles of MME during the procedure. The modifications are focused on the other phases. We mainly details the operations of AuthCtrl and BearerSetup.

MEIA and NAS security setup phases. In traditional LTE/EPC, the functions related to MEIA and NAS security setup phase are located in the MME. We separate these functions from the MME and organize them into the InitialCtrl and SecuCtrl, respectively. The signaling interactions associated with InitialCtrl and SecuCtrl are plotted in Figure 4. In order to maintain the compatibility with traditional LTE/EPC, signaling contexts in NOS-EPC keep unchanged.

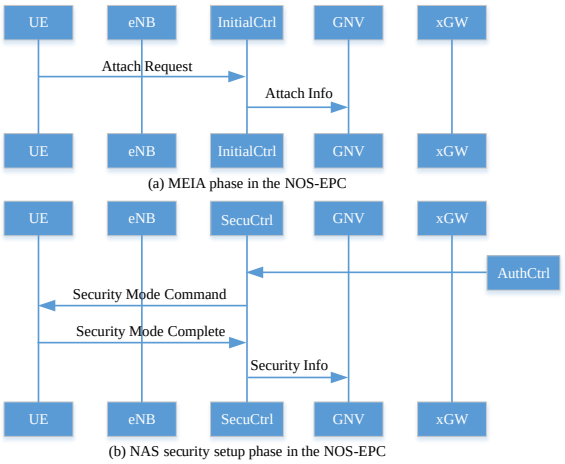


Figure 4: MEIA and NAS security setup phase in the NOS-EPC

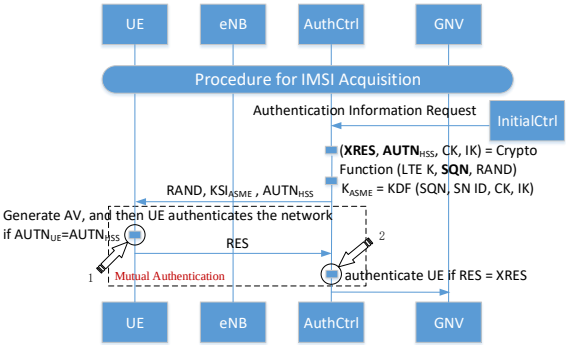


Figure 5: Authentication phase in the NOS-EPC

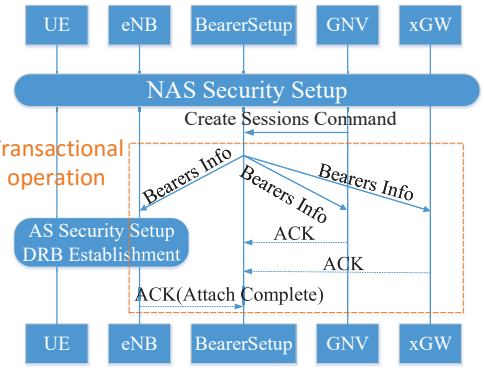


Figure 6: The default bearer setup phase in the NOS-EPC

The authentication phase. In the NOS-EPC, the authentication functions originally dispersed in the MME and HSS are centralized into AuthCtrl. The authentication process in LTE consists of two steps: authentication vector (AV) acquisition and mutual authentication. AV is a quadruple, containing XRES, AUTN, K_{ASME}, RAND. The AuthCtrl generates the AV with two steps, similar to HHS. First, the AuthCtrl calculates XRES, AUTN_{HSS} with a random value RAND.

Next, it uses the key derivation function (KDF) to derive K_{ASME}. During the mutual authentication, UEs authenticate the network with RAND and AUTN_{HSS} (the circle marked with 1 in Figure 5). A RES is then produced if the network has been authenticated (i.e., AUTN_{HSS}=AUTN_{UE}, where AUTN_{UE} is calculated with RAND and AUTN_{HSS}). The network authenticates the UEs by comparing RES and XRES (the circle marked with 2 in Figure 5). In the end, the results are synced to GNV for the records. The details of the signaling interactions for authorization in NOS-EPC are shown in Figure 5.

The location update and default EPS bearer setup. The functions involved in these two phases are originally distributed among HSS, MME, S-GW, P-GW, PCRF, and SPR. In the NOS-EPC, they are reorganized into the BearerSetup. A create session command would be sent by the GNV to the sub-DGCs (BearerSetup). The create session command contains all the necessary information to establish a bearer, including the information in the traditional "update location answer", "profile response" and so on. Once the BearerSetup receives the signaling, it carries out the bearer establishment algorithm, producing the bearer contexts. Then the bearer contexts are disseminated to the U-plane (UE, eNB, xGW) simultaneously. The bearer contexts for different entities are heterogeneous. The signaling interactions involved by BearerSetup in NOS-EPC are shown in Figure 6. The results (bearer context) of the sub-DGCs (BearerSetup) would be updated to the GNV to ensure the completeness.

4.2 Service request

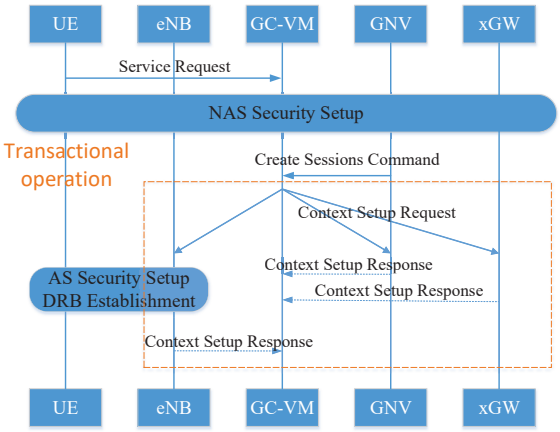


Figure 7: The service request procedure in the NOS-EPC

The service request could be trigged by UEs (uplink transmission) or network (downlink transmission). In this paper, we emphasis the procedure trigged by UEs in Figure 7. The procedure is executed when UEs wants to use a service from the Internet or the PDN networks.

We first focus on the signalling flows during the service request procedure. The main purpose is to setup a data forwarding path between the UE and its destination. The procedure is initiated when a service request message is sent to BearerSetup. Once BearerSetup receives the service request messages, it performs an integrity check on NAS-MAC. If the authentication check is passed, BearerSetup obtains the user-service-related information from GNV, including the UE subscription levels, requested service types, overall network congestions. According to these information, BearerSetup setups the forwarding path by distributing the context setup requests. The context setup requests include the IDs (i.e., TEIDs, bearer IDs), which identify the traffic channels for different UEs among NEs.

The establishment of data forwarding path involves resource allocation, e.g., routing scheduling, QoS mechanism, etc. On the aspect of route scheduling, efforts can be made to improve the capacity of U-plane. Since SDN and NFV technique are used, NOS-EPC is highly programmable. Adaptive routing strategies can be adopted to reduce the congestions by configuring the flow tables. Examples can be found in [24, 25], in which operators jointly optimize task routing and cloud selection to improve the system cost. Meanwhile, the implementation is more convenient in NOS-EPC owing to GNV, which has a global view of the network. BearerSetup can direct the traffic by configuring the flow table in the U-plane, bypassing the congestion links and improving the transmission capacity [26]. Other than routing strategy, advanced resource allocation algorithms or framework like [14, 15] can be migrated into NOS-EPC directly.

Resource allocation is another main concern since QoS mechanisms need to be incorporated to support multiple services with different QoS requirement. To support the legacy of LTE/EPC, conventional QoS mechanisms are fully migrated to NOS-EPC. In order to assign network resources to the traffic of each UE and manage them properly, BearerSetup acquires the subscriptions of UEs and services from GNV through create session command. BearerSetup first classifies user traffic into different service data flows (SDFs). Each SDF contains the flows with the same QoS parameters, e.g., QCI. Different QoS parameters are meant for different priorities of traffic flow which is shown in Table 1. BearerSetup appends QoS rules into Contest setup requests. The enforcement of QoS rules on NEs in the U-plane is the same as the cases in LTE/EPC, which is detailed in [23].

4.3 X2 Handover

4.3.1 Handover in LTE/EPC

X2 handover allows a UE to move from a serving cell to a neighbor cell and reestablish a new radio resource connection(RRC)[27]. Before the handover, the UE collects measurement reports about the signal strength and the quality of adjacent cells, and sends them to the source eNodeB(SeNB). SeNB determines whether the UE initiates handover to a target eNodeB(TeNB) or not. Three phases of handover are successively carried out.

Table 1: The procedure duration in different architectures

QCI	Resource Type	Priority	Packet delay budget	Packet error loss rate	Example services
1	GBR	2	100 ms	10^{-2}	Conversational voice
2		4	150 ms	10^{-3}	Conversational video (live streaming)
3		3	50 ms	10^{-3}	Real time gaming
4		5	300 ms	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100 ms	10^{-6}	IMS signalling
6		6	300 ms	10^{-6}	Video (buffered streaming)
7		7	100 ms	10^{-3}	Video (live streaming)
8		8	300 ms	10^{-6}	Video (buffered streaming)
9		9			

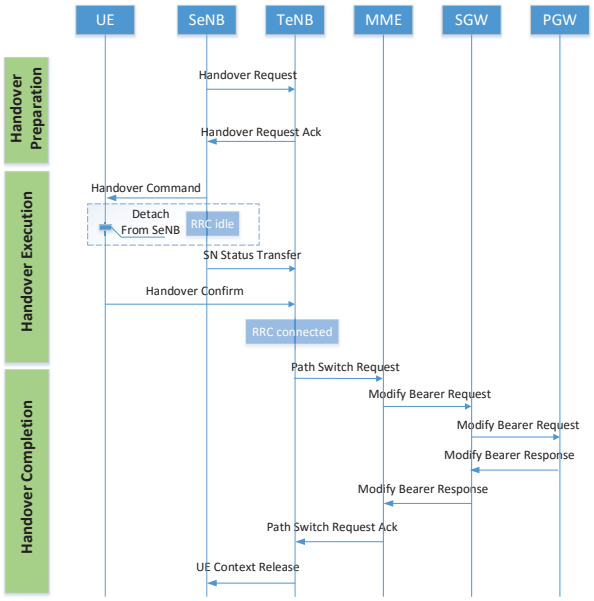


Figure 8: The signaling flow for X2 handover in LTE/EPC

Phase 1: Handover preparation. SeNB forwards the UE context along with S1 TEID to TeNB, assisting with the uplink from TeNB to SGW. The SeNB TEID is sent to TeNB to establish the X2 transport bearer, through which TeNB receives downlink packets while UE attempts to attach TeNB.

Phase 2: Handover execution. SeNB informs the UE to perform handover by sending a message related to TeNB. The message includes cells radio network temporary identity, access stratum security algorithm, *etc*. The UE detaches from SeNB and access TeNB, in the meantime the RRC connection is required to be reconfigured.

Phase 3: Handover completion. When the UE has been connected to the target cell, TeNB is obligated to inform MME to modify the bearer path accordingly. TeNB transmits its TEID to SGW and PGW via MME. The response has to double back. SGW sends the downlink packets to TeNB through the newly established downlink bearer.

Consider the latency during the handover procedure in LTE/EPC. The latency could be presented in (1).

$$D^{\mathrm{LTE}} = 5D_{\mathrm{X2}} + 2D_{\mathrm{RRC}} + 2D_{\mathrm{S1}} + 2D_{\mathrm{S11}} + 2D_{\mathrm{S5/8}} \quad (1)$$

where D_{X2}, D_{RRC}, D_{S1}, D_{S11}, and $D_{\mathrm{S5/8}}$ are the latency motivated by X2, RRC, S1, S11, and S5/8, respectively. It is noticed that (1) assume successful transmission at first attempt. The actual delay values can be higher if some steps require re-transmissions.

4.3.2 Handover in NOS-EPC

In NOS-EPC, the control functions are converged into GC-VMs. We use eNB-U represents eNB that exclusively retains user functions. The handover-related GC-VMs control the handover procedure, which is shown in Figure 9.

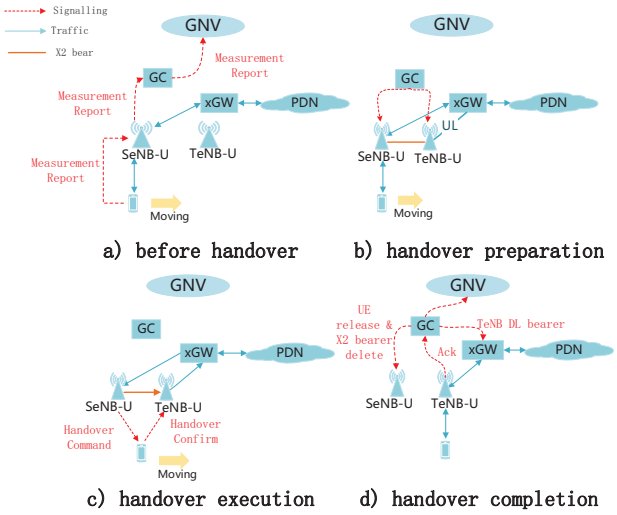


Figure 9: The Handover procedure in NOS-EPC

Initially, an UE reports its measurement to the network. Next, GC-VM accesses users' information from GNV and makes the handover decision. The decision determines when and where to perform the handover procedure. Also, it points out which target cell the UE is about to attach.

Phase 1: Handover preparation. When the TeNB is determined, GC-VM pass the handover execution information to SeNB. Also, GC-VM dispatches TEID of SeNB and xGW to the TeNB. The former TEID is for X2 transport downlink bearer setup. The latter one is for the establishment of an xGW uplink bearer. The setup of X2 transport downlink bearer ensures that the

data could be buffered in target base station during the handover execution.

Phase 2: Handover execution. Once receiving the handover command from GC-VM, the UE obtains the information (e.g., access stratum keys) for attaching to the TeNB. There are no packet transmissions until it successfully accesses to TeNB. Packets from PDN are temporarily stored in the TeNB. The state of non-access stratum layer on UE remains unchanged during the handover execution. The radio link connection is inactive. Then, TeNB connects to the UE with access stratum security algorithms.

Phase 3: Completion. Once the UE connects to the TeNB, it informs the GC-VM that the handover procedure is completed. Then, GC-VM instructs SeNB to release the UE context and delete the previous tunnel (i.e., the tunnel between xGW and SeNB). Meanwhile, GC-VM distributes TeNB's TEID to xGW, which helps to establish the downlink bearer from xGW to TeNB. Also, the instantaneous location and the context of UE are updated to GNV.

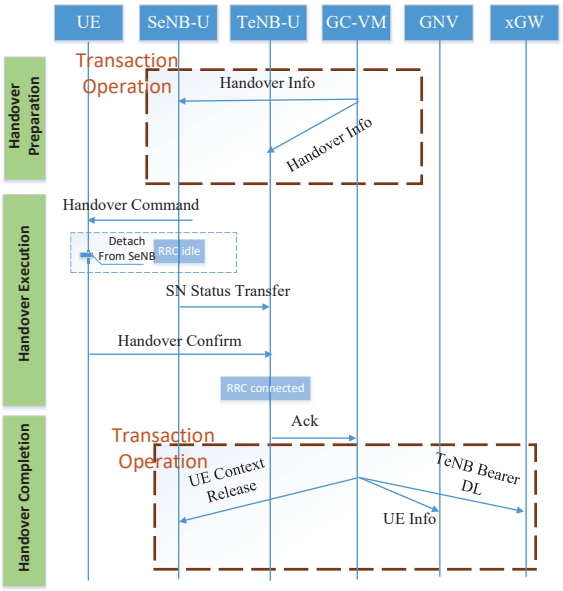


Figure 10: The Signaling flow for X2 handover in NOS-EPC

Consider the latency during the handover procedure in LTE/NOS-EPC. The latency could be presented in (2)

$$D^{\text{NOE-EPC}} = 2D_{\text{TO}} + 2D_{\text{RRC}} + D_{\text{X2}} + D_{\text{CU}} \qquad (2)$$

where D_{RRC} and D_{X2} are the latency motivated by RRC and X2. D_{TO} is the latency incurred by transaction operation. D_{CU} is latency between eNB and GC. The actual delay values can be higher if some steps require re-transmissions.

4.4 detach

In the detach procedure, the UE is detached/detaches from the network he attached to. After this procedure, the network resources associated with the UE are completely released.

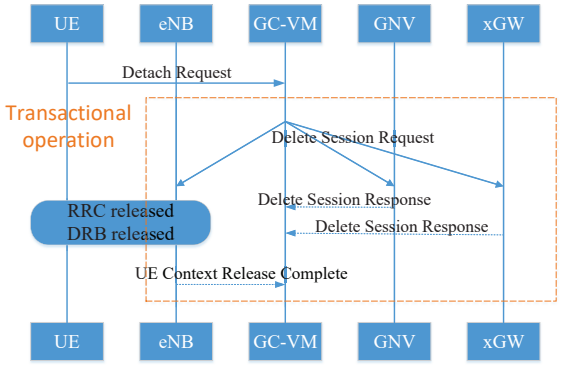


Figure 11: The detach procedure in NOS-EPC

The detach procedure in NOS-EPC is shown in Figure 11. It is an inverse operation of initial attachment. We assume that the detach procedure is triggered by the UE. Initially, the UE sends the detach request to the GC-VMs. The detach request mainly contains the GUTI and detach type of UE. Once the GC receives the detach request, it dispenses the delete session requests to eNB, GNV, and xGW simultaneously. The requests send to different NEs are heterogeneous, e.g., the request sent to eNB contains the 'detach accept' to the UE. These NEs (i.e., eNB, GNV, and xGW) delete the local context about UEs, i.e., network resources and return the response messages to GC-VM.

4.5 The Incorporation of new applications/services in NOS-EPC

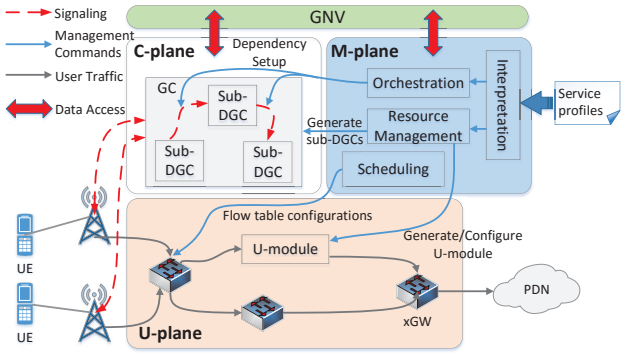


Figure 12: The integration of innovative applications in NOS-EPC

The process of integrating a new application/service in NOS-EPC is shown in Figure 12. First, the service provider submits the profiles into the M-plane through the interpretation module. The profiles contain the functions and the requirements of the new application/service. The interpretation module is essentially a man-machine interface, which analyzes the profiles and generates a series of commands. The control functions are constructed as DGCs/sub-DGCs. The user functions are constructed as U-plane modules. The dependency among modules is configured into NE's flow tables. The meta information (e.g., traffic template) are synced into the GNV.

5 The comparisons between NOS-EPC and the other EPC solutions

In this section, we compare our solutions with the traditional EPC and the SDN-based EPC in order to analyze the pros and cons of our NOS-EPC. Firstly, we brief the architectures of traditional EPC and SDN-based EPC.

Table 2: Comparisons between different architectures

	LTE	SDN-EPC	NOS-EPC
OPEX	High	Med	Low
Scalability	Low	Low	High
Flexility	Low	Med	High
Overheads	Med	High	Low
Capability	Low	High	High

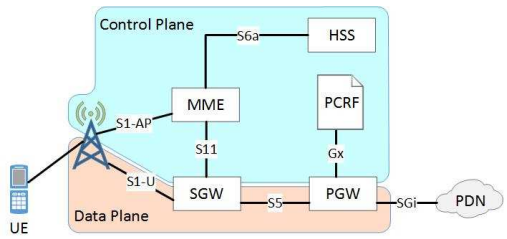


Figure 13: The architecture of the traditional LTE/EPC

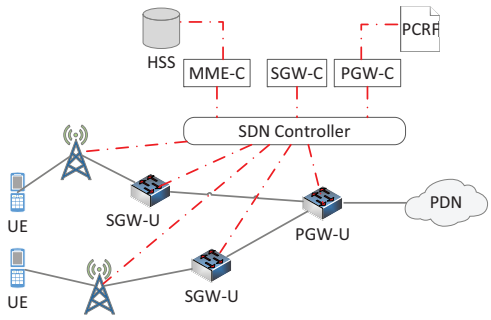


Figure 14: The architecture of the EPC based on SDN

The architecture of traditional LTE/EPC is shown in Figure 13. The details could be referred to [28]. The SDN-based architecture is shown in the Figure 14. In SDN-EPC, the control plane and data plane are decoupled by an SDN controller (i.e. floodlight). NEs (i.e. MME, SGW-C, SGW-P) in the control plane are implemented as SDN applications. The functions in the data plane are achieved with the extended SDN+ switches. The architecture of SDN-EPC could be referred to [29].

Table 2 compares these three architectures, covering OPEX, flexibility, scalability, signaling overheads, and user traffic transmission capability. We detail the comparisons as followings.

OPEX: The NEs (i.e. MME, SGW, PGW, HSS, SPR) in the traditional EPC are deployed on dedicated hardware. The special-designed hardware for different NEs increases the cost of procurement. The NOS-EPC and the control plane of SDN-EPC apply the virtualization technique. Commercial servers are used, which means the OPEX could be reduced.

Flexibility: Flexibility refers to the ability of adaption. In the traditional EPC, any modifications require considerations for the entire system because of the distributed control. The control paradigms in SDN-EPC are not changed compared with traditional LTE network. Therefore, the SDN-EPC has the exact same problems. On the other hand, in our NOS-EPC architecture, innovative services could be easily incorporated, as we described earlier. Thus, the flexibility of the proposed NOS-EPC is more superior.

Scalability: In order to improve the capabilities of the traditional EPC, operators usually need to replace the NEs with more powerful appliances. The replacements need to be accomplished locally and manually. Moreover, the traditional EPC typically scale vertically [29], which means the enhancements could not target at specific services. SDN-EPCs have similar issues. However, it would be easier in the SDN-EPC due to virtualization. Unfortunately, the SDN-EPC introduces bottleneck nodes (SDN controllers) between the control plane and the user plane. The scalability would be largely affected by the capacity of bottleneck nodes. Although some advanced techniques are proposed to deploy SDN in a large-scale network, there are still some problems need to be solved so far [30].

In the NOS-EPC, the 'logical centralized' GNV can be deployed as distributed databases, which has been widely used in the commercial market [21]. Supported by GNV, it is quite convenient to enhance system capacity by deploying replicas of modules and DGCs. Moreover, system enhancements can be accomplished remotely and automatically due to the existence of M-plane and GNV. Horizontally scaling of the EPC could be achieved through the duplications of the related DGCs/sub-DGCs or U/M-modules. The access loads on GNV could be relieved by distributed deployment (distributed databases).

Distinct from the SDN-base architecture, in NOS-EPC, the capacity of the connections between different planes (U/C/M plane) would not be limited by a single node (or host/server). The diagnosis modules and the scheduling modules would automatically route the traffic to bypass the congestion links.

Signaling overheads: The distributed control in the traditional EPC or SDN-EPC contributes the coordination signaling and complicated handshakes among NEs. Additionally, in SDN-EPC, more signaling and handshakes are needed due to the separations of PGW and SGW. On the other hand, in our proposed NOS-EPC, the logic for a particular LTE procedure or service is encapsulated into DGCs/sub-DGCs. The signaling overheads introduced by cooperation and handshakes are cutoff. The performance of the NOS-EPC would be considerably improved.

Moreover, in LTE, the signaling is delivered layer by layer. For example in the handover procedure, the signaling is successively enforced to MME, SGW and PGW when TeNB establishes bearers between

SGW/PGW. However, in the NOS-EPC, signaling interactions among NEs is no longer constrained by layers. Thus, signaling overheads and handshakes are significantly cut down. In fact, LTE/EPC requires 12 signaling interactions during the handover. However, in NOS-EPC, only 9 interactions are required.

Transmission capability: In the EPC, packet loss is typically unacceptable. In traditional EPC, more resources (i.e. bandwidth, links, etc.) need to be provisioned to ensure the reliability. In dynamic network surroundings, performance degradation due to congestion is inevitable even with more resources. In SDN-EPC and NOS-EPC, the traffic routes could be optimized globally. A proper control can improve the average link utilization from 30%∼40% to nearly 100% [31].

There are some costs for our NOS-EPC as well. The replacements of dedicated hardware with commodity servers lead to consistent degradation (nearly 7% degradation [32]) on the aspect of the processing capacity. Second, the deployments of GNV need to be further investigated.

6 Performance evaluations

In this section, we evaluate the performance of the NOS-EPC. We mainly concentrate on the performance of C-plane in different architectures (traditional EPC/EPC, SDN-EPC, and NOS-EPC). We compare the network efficiency on procedure durations (i.e., latency) and signaling overheads. The comparisons highlight the superiority of NOS-EPC on the aspects of procedure performance and network efficiency. The results also suggest that the NOS can help to build a promising cellular network. The comparisons are conducted with NS3.

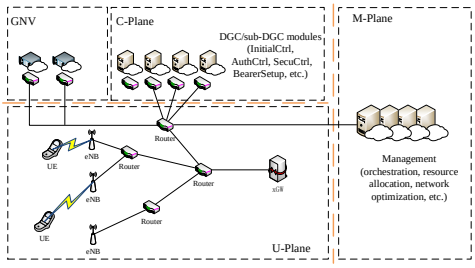


Figure 15: Simulation setting for NOS-EPC.

In the simulations, we use Cost266 as the network topology of NOS-EPC [33]. We assume that nodes in Cost266 are equipped with computation resources. We deploy sub-DGCs/DGC (e.g., InitialCtrl, SecuCtrl, Authctrl, and BearerSetup), U/M modules, and GNV as NS3 applications on separate nodes. eNBs are added which connects the nodes in Cost266. We deploy eNBs in a line with different cell size (R_{cell}), connecting the nodes in Cost266. A complete NOS-EPC is then built as Figure 15. The bandwidth of links between the nodes is uniformly configured as 1Gbps.

UEs in Figure 15 are also implemented with nodes in NS3. In this paper, we mainly focus on the latency and signaling overheads motivated by initial attachment, service request, handover, and detach procedure. Thus, we set that each UE attaches to the EPC with an average interval equalling 250 ms. Once an initial attachment procedure is triggered, the following procedures (service request, detach) are sequentially invoked. The handover procedure is initialized when UE moves out of the current eNB's coverage. We change the number of UEs to vary workloads of EPC. In the simulations, UEs move on a line with different speed (V_{UE}). The parameters for the simulations are summarized in Table 3.

We carry out the simulations in both static scenarios (without UE mobility) and dynamic scenarios (with UE mobility). In static scenarios, we rule out the influences brought by UE mobility and address the performance of initial attachment, service request, and detach. In dynamic scenarios, we focus on the influences under different cell size and UE velocity.

We first evaluate the performance in a static environment ($V_{\text{UE}} = 0$). Table 4 shows the procedure duration in different architectures. The percentages in parentheses tell the ratios using the traditional EPC as the baseline. The variances of the duration in the table indicate the jitters of the procedure duration. From the points of average duration and jitters, the NOS-EPC takes advantages. Compared with the traditional EPC, the NOS-EPC could decrease the procedure duration by nearly 42.5%. On the other hand, the SDN-EPC shows its degradation about 27.5%. The average procedure duration is mainly affected by the number of interactions among different NEs. The number of interactions in different architectures is shown in Table 5. In this table, each transaction operation is viewed as an interaction. It could be seen that NOS-EPC decreases the signaling overheads by at least 25% compared to LTE/EPC. On the other hand, the SDN-EPC suffers from redundant signaling overheads due to its crude U/M decoupling.

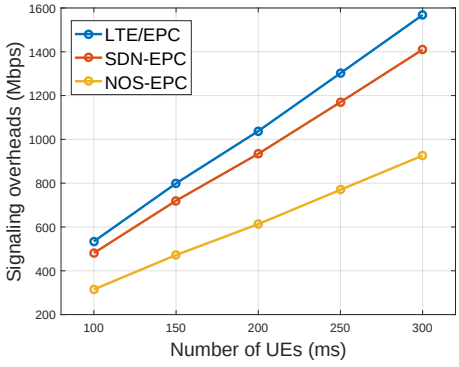


Figure 16: The signaling overheads in different architectures.

Figure 16 shows the signaling overheads in different architectures, which coincides with the conclusions implied by Table 5. It proves that our NOS-EPC could sufficiently decrease the signaling overheads. The reductions in NOS-EPC could be as much as 35% com-

Table 3: Simulation parameters

Description	Values
UE number	100~300
Average interval of initial attachment	25 ms
UE velocity (V_{UE})	0 ~ 40 m/s
Cell size (R_{cell})	150 ~ 400 m
Link bandwidth	1 Gbps

Table 4: The procedure duration in different architectures

	Number of users	Average duration (ns)	Variance
LTE EPC	100	160047.06(=1)	0.25
	150	160047.10(=1)	0.40
	200	160047.12(=1)	0.45
	250	160047.15(=1)	0.59
	300	160047.19(=1)	0.77
SDN-EPC	100	204055.07(~+27.50%)	0.26
	150	204055.11(~+27.50%)	0.50
	200	204055.15(~+27.50%)	0.71
	250	204055.17(~+27.50%)	0.79
	300	204055.21(~+27.50%)	0.97
NOS-EPC (Proposed)	100	92028.02 (~-42.50%)	0.08
	150	92028.03 (~-42.50%)	0.11
	200	92028.04 (~-42.50%)	0.15
	250	92028.05 (~-42.50%)	0.18
	300	92028.07 (~-42.50%)	0.24

Table 5: The number of interactions in different architectures

	Initial attach	Detach	Service request	Handover
LTE EPC	19	6	6	12
SDN EPC	23	10	10	16
NOS-EPC	10	3	3	6

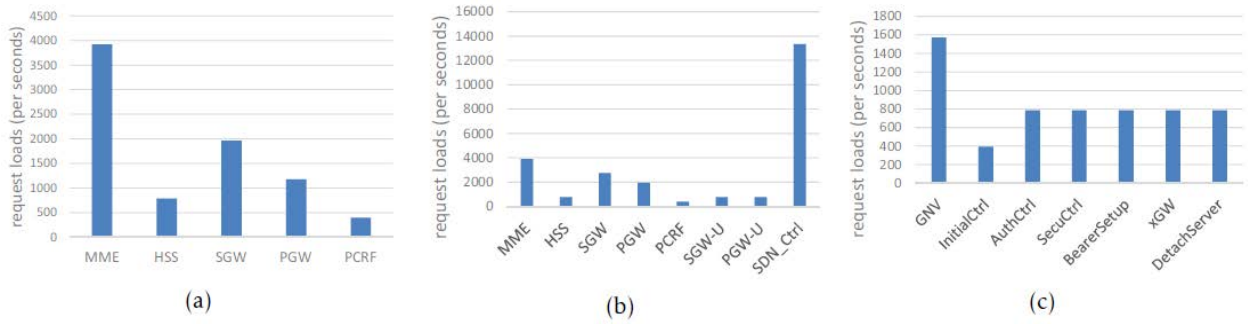


Figure 17: The request loads on the NEs in different architectures. The loads are measured in the simulations with 10 UEs. (a) shows the loads on the NEs in traditional EPC. (b) shows the loads of the NEs in SDN-EPC including the SDN controllers. (c) shows the cases in NOS-EPC. The strings in x label denote the name of sub-DGCs in the C-plane.

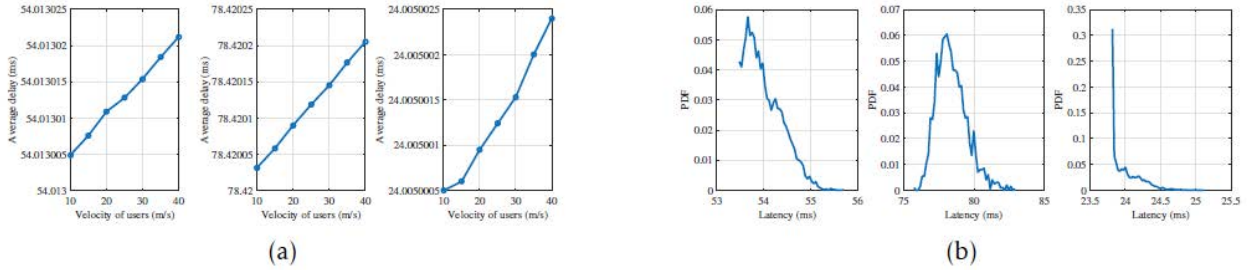


Figure 18: The handover latency in different architectures. (a) The handover latency with different UE velocity when $R_{\text{cell}} = 150m$. (b) The probability distribution function (PDF) of handover latency with $V_{\text{UE}} = 40m/s$ and $R_{\text{cell}} = 150m$.

pared with traditional EPC. The signaling overheads also influences the jitters. The queue delays (as random variables) become significant as the overheads increases. Therefore, the jitters of transmissions become larger.

Figure 17 shows the request loads for different NEs in three architectures with $V_{\text{UE}} = 0$. In traditional EPC, the pressure on the MME is the heaviest. In the SDN-EPC, the bottleneck node (SDN controller) has the highest request loads. Other than GNV, the distribution of request loads in NOS-EPC is nearly balanced. The heavy loads on GNV could be relieved if the GNV is deployed in a distributed manner. The requests loads on different NEs imply the scalability of different EPCs. The SDN controller and MME would become the bottlenecks in SDN-EPC and traditional EPC, respectively. In the NOS-EPC, things become less tricky due to virtualization and nature of the architecture. The distributions of loads on different NEs also affect the duration. A congested path would largely increase the variance of the latency [34]. With a heavy loaded SDN controller, the jitters in NOS-EPC are more significant. The results are agreed with prior results (jitters) in Table 4.

In order to emphasis NOS-EPC performance in a dynamic environment, we evaluate the performance of NOS-EPC. We vary V_{UE} and R_{cell} in the range of $5 \sim 40$m/s and $150 \sim 400$m, respectively.

Figure 18(a) plots the tendency of handover latency with different UE velocity. It shows that the average handover delay stays about 54.01ms, 78.42ms, and 24.00ms in LTE/EPC, SDN-EPC, and NOS-EPC, respectively. The delay slightly increases with the UE velocity since a higher V_{UE} results in higher congestion in EPC. The proposed architecture (NOS-EPC) reduces the delay by nearly 55.5% compared to the LTE/EPC. The benefit is owing to the redesigns of the control paradigm. Redundant signaling interactions are cut off due to the centralized control. On the other hand, SDN-EPC has the longest handover delay. The reason is that more signaling overheads are required due to the sync between SGW(or PGW) and SGW-U(or PGW-U).

Figure 18(b) compares the latency distributions of handover procedure in different architectures. The left graph is for the handover procedure in LTE/EPC. The middle one represents the case in SDN-EPC and the right graph is the case in NOS-EPC. The shapes of the curves in Figure 18(b) tells the probability distributions of the handover procedure. Moreover, the peaks of the curves describe the number of network backlogs. When the network suffers severe congestion (backlogs), the queue delay on NEs increases, shifting the peaks to the rightwards. Vice Versa. Therefore, the procedure in NOS-EPC causes fewer network backlogs than the other network architectures.

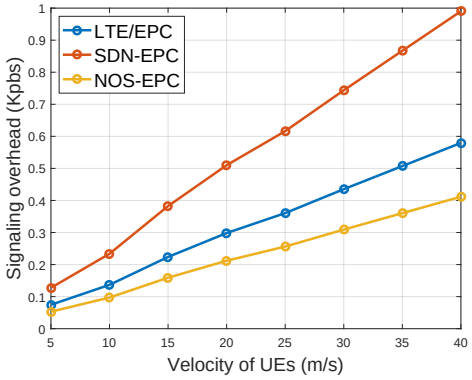


Figure 19: The signaling overheads brought by a single UE with different V_{UE}. We set $R_{\text{cell}} = 150m$.

Figure 19 shows the relationships between the signaling overheads and V_{UE}. As V_{UE} increases, the handover procedure is invoked more frequently, which increases the signaling overheads. Compared with LTE/EPC, the proposed NOS-EPC could always cut off the signaling overheads by nearly 28.8%. In NOS-EPC, the overall information elements in handover signaling decrease from 373 bytes to 265 bytes. The SDN-EPC take more signaling overheads compared to LTE/EPC and SDN-EPC since it requires additional signaling to sync the U/C plane, e.g., PGW and PGW-U, SGW and SGW-U.

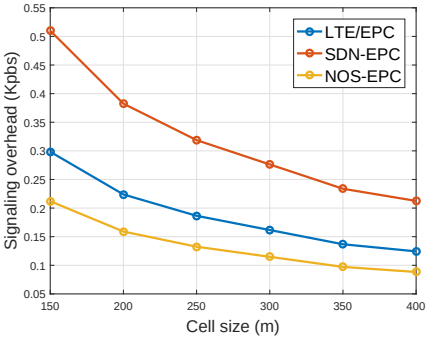


Figure 20: The signaling overheads brought by a single UE with different R_{cell}. We set $V_{\text{UE}} = 10m/s$.

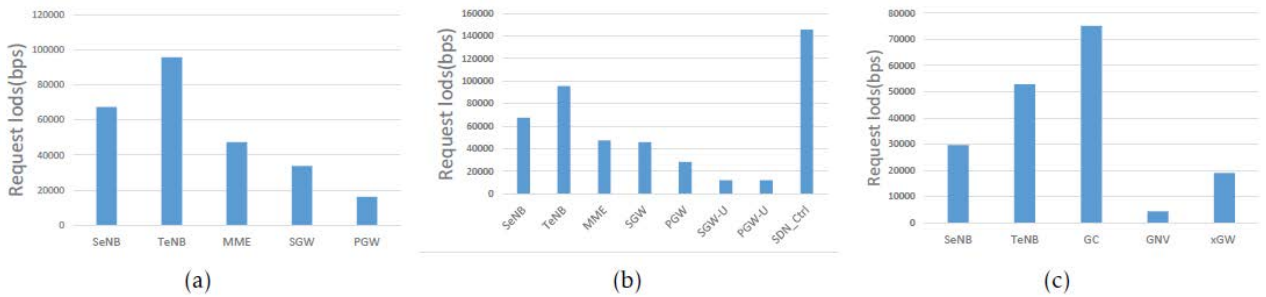


Figure 21: The request load brought by handover in different architectures. The loads are measured in the simulations with 2000 UEs. (a) shows the loads on the NEs in traditional EPC. (b) shows the loads on the NEs in SDN-EPC including the SDN controllers. (c) shows the cases in NOS-EPC.

Figure 20 plots the relationships between the signaling overheads and R_{cell}. A small-size cell requires frequent handovers. As R_{cell} increases from 200m to 400m, the handover times decrease by nearly 54.4%. Therefore, when R_{cell} increases, signaling overheads decreases. Coincided with former analysis, the overheads in NOS-EPC decrease by nearly 28.8% compared to LTE/EPC. Meanwhile, SDN-EPC performs the worst owing to a roughly decouple between U-plane and C-plane.

Figure 21 shows the overheads on NEs in different architectures. Similar to the static scenarios, the proposed NOS-EPC shows relatively balanced distributions. In LTE/EPC and SDN-EPC, MME and SDN controller are the bottlenecks which would significantly degenerate the system performance. The heavy load on these NEs would result in additional network cost, such as packet loss, workload expiration, etc. On the other hand, the load in NOS-EPC is nearly balanced. Moreover, supported by GNV, the capacity of NEs could be easily enhanced by the duplication of NE instances.

7 Conclusion

In this paper, we proposed an innovative EPC architecture based on NOS framework, which consists of the U-plane, C-plane, M-plane, and GNV. We describe the operation of the NOS-EPC with four common EPS mobility management(EMM) procedures, i.e., initial attachment, service request, handover and detach. We detail the comparison among conventional LTE/EPC, SDN-EPC, and the proposed NOS-EPC on OPEX, scalability, flexibility, signaling overheads, and transmission capability. Compared to the conventional LTE/EPC, the proposed NOS-EPC takes advantages since we loosely decouple the U/C/M plane through GNV. C/M plane is carefully designed to deliver a centralized control. The GNV is deployed in a distributed manner. Thus, no bottleneck node exists in the proposed architecture compared to the SDN-EPC. The simulations verify that compared with LTE/EPC and SDN-EPC, NOS-PEC could significantly decrease the signaling overheads (by at least 35%) and procedure duration (by at least 42%). Moreover, we conduct the simulations to evaluate the workload on different NEs. The results show that the load distribution in NOS-EPC is nearly balanced. On the other hand, LTE/EPC and SDN-EPC suffer from severe congestion on MME and SDN controller, which decreases the performance on scalability. The advantages on the scalability, programmability and the improvements of the procedure duration, signaling overheads, and loads distribution mean that the NOS-EPC is a promising solution for the B5G network. In the future work, we will focus on the division granularity of the GC modules and GNV, which can be further optimized to enhance the performances of the network. Moreover, a testbed for NOS-EPC evaluation would be developed.

Acknowledgment This work is supported by the National S&T Major Project(No.2018ZX03001004).

References

[1] S. Sekander, H. Tabassum, and E. Hossain, "Multi-Tier Drone Architecture for 5G/B5G Cellular Networks: Challenges, Trends, and Prospects," IEEE Communications Magazine, 56(3), 96-103, March 2018.

[2] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," IEEE Communications Magazine, 54(4), 84-91, April 2016.

[3] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless Network Virtualization with SDN and C-RAN for 5G Networks: Requirements, Opportunities, and Challenges," IEEE Access, 19099-19115, May 2017.

[4] Q. Duan, N. Ansari, and M. Toy, "Software-defined network virtualization: an architectural framework for integrating SDN and NFV for service provisioning in future networks," IEEE Network, 30(5), 10-16, Sep. 2016.

[5] B. Wu, L. Ge, J. Zeng, X. Zheng, Y. Kuang, X. Su, and J. Wang, "An Innovative EPC with Not Only Stack for beyond 5G Mobile Networks," 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), 1-5, June 2018.

[6] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," IEEE Communications Magazine, 51(7), 44-53, Jul. 2013.

[7] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "CONCERT: a cloud-based architecture for next-generation cellular systems," IEEE Wireless Communications, 21(6), 14-22, Dec. 2014.

[8] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," Computer Networks, 85, 1-18, 2015.

[9] X. Sun and N. Ansari, "Green cloudlet network: A distributed green mobile cloud network," IEEE Network, 31(1), 64-70, Jan. 2017.

[10] S. Shanmugalingam and P. Bertin, "Programmable Mobile Core Network," 2014 IEEE Symposium on Computers and Communications (ISCC), 1-7, June 2014.

[11] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "EASE: EPC as a service to ease mobile core network deployment over cloud," IEEE Network, 29(2), 78-88, March 2015.

[12] M. R. Sama, S. B. H. Said, K. Guillouard, and L. Suciu, "Enabling network programmability in LTE/EPC architecture using OpenFlow," 2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 389-396, May 2014.

[13] V. Nguyen and Y. Kim, "Proposal and evaluation of SDN based mobile packet core networks," EURASIP Journal on Wireless Communications and Networking,2015(1), 172-190, 2015.

[14] V. D'Amico, A. Lombardo, M. Melita, C. Rametta, and G. Schembra, "An SDN/NFV telco operator platform for video broadcasting," IEEE Communications Magazine, 54(10), 122-128, Oct. 2016.

[15] J. Costa-Requena, A. Poutanen, S. Vural, G. Kamel, C. Clark, and S. K. Roy, "SDN-Based UPF for Mobile Backhaul Network Slicing," 2018 European Conference on Networks and Communications (EuCNC), 48-53, June 2018.

[16] X. An, W. Kiess, J. Varga, J. Prade, H. Morper, and K. Hoffmann, "SDN-based vs. software-only EPC gateways: A cost analysis," 2016 IEEE NetSoft Conference and Workshops (NetSoft), 146-150, June 2016.

[17] E. B. Hamza and S. Kimura, "A Scalable SDN-EPC Architecture Based on OpenFlow-Enabled Switches to Support Inter-domain Handover," 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 272-277, July 2016.

[18] S. Zhou, T. Zhao, Z. Niu, and S. Zhou, "Software-defined hypercellular architecture for green and elastic wireless access," IEEE Communications Magazine, 54(1), 12-19, Jan. 2016.

[19] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "5G virtualized radio access network approach based on NO Stack framework," Communications (ICC), 2017 IEEE International Conference on, 1-5, IEEE, 2017.

[20] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E. Schmidt, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," 2013 IEEE SDN for Future Networks and Services (SDN4FNS), 1-7, Nov. 2013.

[21] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: Amazon's Highly Available Key-value Store," SIGOPS Oper. Syst. Rev., 41(6), 205-220, Oct. 2007.

[22] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J. J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, W. Hsieh, S. Kanthak, E. Kogan, H. Li, A. Lloyd, S. Melnik, D. Mwaura, D. Nagle, S. Quinlan, R. Rao, L. Rolig, Y. Saito, M. Szymaniak, C. Taylor, R. Wang, and D. Woodford, "Spanner: Google&Rsquo;s Globally Distributed Database," ACM Trans. Comput. Syst., 31(2), 8:1-8:22, Aug. 2013.

[23] Netmanias, "EMM Procedure 1. Initial Attach - Part 2. Call Flow of Initial Attach." https://www.netmanias.com/en/?m=view&id=techdocs&no=6102/. Accessed Jan. 16, 2014.

[24] A. Ceselli, M. Premoli, and S. Secci, "Mobile Edge Cloud Network Design Optimization," IEEE/ACM Trans. Netw., 25(3), 1818-1831, June 2017.

[25] B. Yang, W. K. Chai, Z. Xu, K. V. Katsaros, and G. Pavlou, "Cost-Efficient NFV-Enabled Mobile Edge-Cloud for Low Latency Mobile Applications," IEEE Transactions on Network and Service Management, 15(1), 475-488, March 2018.

[26] J. Hansen, D. E. Lucani, J. Krigslund, M. Medard, and F. H. P. Fitzek, "Network coded software defined networking: enabling 5G transmission and storage networks," IEEE Communications Magazine, 53(9), 100-107, Sep. 2015.

[27] 3GPP TS 36.420 V14.0.1, "X2 general aspects and principles (Release 14)," 1, 0-14, 2017.

[28] U. Barth, "3GPP Long-Term evolution/system architecture evolution overview," Alcatel White Paper, 2006.

[29] A. Jain, N. Sadagopan, S. K. Lohani, and M. Vutukuru, "A comparison of SDN and NFV for re-designing the LTE packet core," Network Function Virtualization and Software Defined Networks (NFV-SDN), IEEE Conference on, 74-80, IEEE, 2016.

[30] T. Huang, F. R. Yu, C. Zhang, J. Liu, J. Zhang, and Y. Liu, "A Survey on Large-Scale Software Defined Networking (SDN) Testbeds: Approaches and Challenges," IEEE Communications Surveys Tutorials, 19(2), 891-917, Secondquarter 2017.

[31] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat, "B4: Experience with a globally-deployed software defined WAN," SIGCOMM Comput. Commun. Rev., 43(4), 3-14, Aug. 2013.

[32] D. Lake, G. Foster, S. Vural, Y. Rahulan, B. H. Oh, N. Wang, and R. Tafazolli, "Virtualising and orchestrating a 5G evolved packet core network," 2017 IEEE Conference on Network Softwarization (NetSoft), 1-5, Jul. 2017.

[33] S. Orlowski, R. Wess?ly, M. Piro, and A. Tomaszewski, "SNDlib 1.0 Survivable Network Design Library," Networks, 55(3), 276-286.

[34] D. E. Comer, Computer networks and Internets. Prentice Hall Press, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 329-342 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Managing and Optimizing Quality of Service in 5G Environments Across the Complete SLA Lifecycle

Evgenia Kapassa[*,1], Marios Touloupou[1], Panagiotis Stavrianos[1], Georgios Xylouris[2], Dimosthenis Kyriazis[1]

[1]*University of Piraeus, Department of Digital Systems, Piraeus, Greece*

[2]*Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece*

A R T I C L E I N F O

Article history:
Received: 20 December, 2018
Accepted: 15 February, 2019
Online : 28 February, 2019

Keywords:
Quality of service provisioning
5G networks
Service level agreements
Cost-efficient monitoring
Network service recommendation

A B S T R A C T

The 5G is the fifth generation of mobile broadband, cellular technologies, and networks that promises a major change in mobility by evolving connected business realities. In such an emerging environment, reliable Service Level Agreements (SLA) and anticipation of breaches of Service Level Objectives (SLO) become compulsory. Thus, guaranteeing the required service quality, while also ensuring efficient recourse allocation becomes a challenge. In addition, 5G networks are expected to provide diverse Quality of Service (QoS) guarantees for a wide range of services, applications and users with a variety of requirements. However, there is an increased difficulty in translating user-friendly business terms into resource-specific monitoring attributes that can be used to manage resources in the 5G core network. To address these gaps, an SLA management framework, enabling QoS provisioning is introduced. The aforementioned framework will be supported by an adaptive monitoring algorithm, which removes the static time interval used in the monitoring system, in order to provide highly accurate information in real time, without the produce of unnecessary traffic to the network. The proposed architecture also incorporates a recommendation mechanism to determine the significance of various QoS parameters in order to ensure that relevant QoS metrics are included in the SLAs, using enriched metadata information from a Network Function Virtualization (NFV) Catalogue.

1. Introduction

The explosive growth in mobile data, forces network operators to transform their networks [1]. The emerging 5G in combination with Software Defined Networks (SDN) aims to comply with different quality of service (QoS) requirements in different application scenarios [2]. Nevertheless, promoting delay-limited QoS over emerging 5G wireless networks with limited resources for bandwidth-intensive and time-sensitive scenarios, presents many new challenges not arised in 4G networks [3,4]. Therefore, it becomes crucial to adopt high performance Virtual Network Functions (VNF) and Network Services (NS) that require a plethora of system resources [5,6]. However, this kind of provisioning in a virtualized environment is a challenging task. Various virtual services in fields like Augmented Reality (AR), Virtual Reality (VR), or autonomous guided vehicles (AGVs), require corresponding Service Level Agreements (SLAs) that capture different levels of guarantees as QoS requirements, such as performance and availability. As stated by many Telco's and cloud service providers, they are tempted by the promise of QoS such as high speed/performance, high reliability, low latency, increased capacity, availability and connectivity, as well as dynamic bandwidth allocation from 5G RANs to core networks [7]. As a result, the responsibility to provision the necessary infrastructure recourses and QoS assurances, lies with the 5G network operators and network service providers, in order to ensure this kind of demands. Consequently, there are challenges to address this topic which include a pro-active SLA management framework, an efficient and high adaptive monitoring system on top of the virtualized 5G infrastructure, while also support the provisioning of appropriate SLAs, in terms of business aspects and recourses allocation. One of the most significant challenges, is the role of an SLA Management Framework on top of the 5G Core network [8]. It should be pointed, that managing a virtualized network attempts to preserve acceptable quality, and at the same time push the customers to negotiate with the service providers precise QoS thresholds. Considering that the customer-related QoS requirements are indicated in an SLA, the thing that is vital for the service providers, refers to the assessment of the necessary resources, for each NS that is going to be deployed.

[*]Evgenia Kapassa, University of Piraeus, +30 2104142746 , ekapassa@unipi.gr

To this end, monitoring the infrastructure but also the running services, should be considered as the mediator for supporting QoS provisioning. In 5G networks, where QoS assurance through SLA enforcement is a crucial process, data monitoring is required to evaluate the health of the network [9]. Additionally, service provisioning and assurance tools in NFV should monitor a large number of end points [10]. Keeping that in mind, a frequency analyzer tool could be the key, where real time data and low running costs are essential in this kind of networks. Thus, monitoring and evaluation of monitoring data are considered as important aspects to track the overall progress and reliability of a running service. However, decisions on adaptive and pro-active tracking should be made without the possible absence of information that could negatively impact strategic decisions [11]. Such solutions should also be able to adapt the virtual service during runtime, while at the same time maintain all health and performance historical data as the service evolves [13].

At this point, questions such as how the QoS parameters would be described, or where the SLAs would be stored, and how the service providers could manage all this information arise. Furthermore, the efficient description of the QoS parameters comprises an issue of paramount importance, along with the storage and the management of the SLA descriptors from the service providers. Primarily, for fulfilling this demand in 5G infrastructures, the concept of the NFV Catalogues is introduced. Here comes the vision of the NFV Catalogues, going beyond a plain data store to the promotion of the service offerings and the facilitation of the commercial activity and fluent interaction among the different business stakeholders. The above-mentioned concept not only enables the main storage of the 5G infrastructure, but also, promotes the management and the exploitation of this information. As a result, recommendations mechanisms, or even predicting tools can be developed, providing the most suitable SLAs for the selected by the customers network service [12].

Considering the aforementioned set of challenges, in this paper, we are trying to portray an "SLA-Oriented Framework" which targets the profitable provisioning of QoS guarantees, in 5G environments. The proposed approach is based on mapping the high-level customers' requirements to low-level recourse attributes, as shortly described in [14,15]. The proposed Framework adopts an Artificial Neural Network (ANN) approach, making it easily adaptable to different software components like VNFs and NSs. It is widely known that ANN have many benefits, but the main reason for selecting them is their fault tolerance. This kind of advantage is very important in the SDN environment, where networks are scaled across multiple machines and multiple servers. Additionally, we also introduce an extra set of mechanisms for suggesting the most important QoS parameters, for endorsing the most reasonable and profitable SLAs for diverse services and customers. Furthermore, we also introduce an adaptable "Monitoring 'Framework, in order to supervise the established SLAs. The latter is based on a scheduling algorithm that provides decision logic on probe level as initially described in [17]. On top of them, an enhanced metadata "NFV-Catalogue" is adopted for valuable management of the NSs and the corresponding SLAs.

The remaining of the paper is organized as follows. Section 2 presents the related work and motivation of this work. Section 3 introduces the overall "SLA-Oriented Framework", while in Section 4, 5 and 6 we describe in detail the three major architectural components. Section 7 states the evaluation of the proposed approach in a real 5G testbed, through an end-to-end scenario. Finally, in Section 8, we close up with ideas for future experiments and current study capabilities

2. Background and Motivation

2.1. SLAs & QoS assurance

It is doubtless that Active and advanced work on SLA management for the cloud an 5G infrastructure has been carried out. Therefore, there are many researches for solutions that handle QoS parameters and monitor in an efficient way the guaranteed SLOs. A framework is discussed in [16] to utilize QoS into grid applications. In this paper a performance model is used to calculate the response time and the pricing model to determine the cost of conducting a job. It should be noted that a baseline for our work, is presented in [19], where the architecture of an SLA management system is described. Another intriguing work is described in [20], which demonstrates the importance of the implementation of ANNs in the service- oriented field. In this approach, the ANN's main objective is to set technical goals in terms of quality goals for the design attributes of web service systems. Moreover, the LoM2HiS framework is presented in [21]. In this case, the authors provide a paradigm that implements the reverse process of the one we propose, where low- level specifications are translated into high- level requirements that are used in cloud SLAs. More details concerning the mapping of requirements can be found in [22,23]. Moving forward, to the next generation of networks, the 5G Network Slice Broker [24] is an innovative network element that builds on the capacity broker function block, for advanced RAN sharing considered by 3GPP. It maps incoming SLA requirements to physical resources in connection with network slice requests, having as a result to get a "slice" of the relevant elements of the Radio Access Network (RAN) [25].

In addition, a lot of work has been done to provide guaranteed QoS for enhanced user experience. There was a migration from QoS management at the user equipment level, to QoS management at the network level during the evolution of the QoS management mechanism in 3GPP networks, a shift which maintained also in 5G networks [26]. The QoS level provided in 5G systems should meet the requirements of future Internet industry and go beyond what can today be accomplished with any wireless communication technology. Essential QoS requirements in currently studied 5G systems include: a) maximum acceptable end-to-end latency (delay) less than 5ms and b) reliability around 10-9% or 99.999% [27, 28].

It is worth mentioning, that delay- limited QoS requirements are relatively difficult to guarantee, due to highly differing wireless channels. Alternatively, the statistical delay- limited QoS supply theory has been initially proposed and demonstrated to be a useful method for characterizing and implementing the delay- limited QoS guarantee for wireless real- time traffic [29]. In addition, several works have proposed solutions in the framework of the QoS scheduling [30], including multi-QoS scheduling as investigated in [31,32]. Instead, authors in [33,34] contended that existing QoS mechanisms do not endorse the implementation of specific policies for a group of network users.

It should be pointed out, that a preliminary description of our work can be found in [14,15,18], where the authors presented an approach for mapping the high-level end-user requirements the low-level policy parameters, and at the same time proposed a mechanism for suggesting the most important QoS parameters to the Service/Infrastructure Provider, in order to achieve better QoS assurance. In the present paper, we are going beyond an isolated SLA management framework, by trying to support business guarantees in the overall lifecycle of NSs. To do so, we co-operate the proposed "SLA Framework" with an "NFV-Catalogue", providing recommendations based on benchmarking results, as well as with an advanced "Monitoring Framework" for efficient SLA violations detection.

2.2. Efficient Monitoring

The need for network monitoring is a key enabler for efficient network management. Several works address this domain. As stated in [35], OpenTM was presented, where the integrated features provided in the OpenFlow switches are used to directly and accurately measure the low overhead traffic matrix. OpenTM also utilizes the routing knowledge acquired from the OpenFlow controller to intelligently select the switches from which flow statistics can be obtained, decreasing the load on switching elements. Furthermore, the authors in [36] presented Flow Sense, a push-based approach to performance monitoring in flow-based networks, where they let the network inform regarding performance changes, rather than query for metrics information on demand. The key point is that control messages sent by switches to the controller contain information that allows performance estimation. Moreover, another well-presented monitoring framework for SDNs has been introduced in [37]. The authors propose a software-defined traffic measurement architecture that distinguishes the data plane from the control plane, namely OpenSketch. OpenSketch offers a simple three-stage pipeline (hash, filter and count) in the data plane that can be enforced with commodity switch components and support many measuring tasks. In the control plane, OpenSketch provides a measurement library that automatically configures the pipeline and allocates resources for different measurement tasks. What is more, in [38], authors have presented an extension of Prometheus.io, a monitoring framework implemented and integrated within SONATA project [39]. In short, the SONATA monitoring framework gathers and processes data from many sources, enabling the developer to activate measurements and thresholds to capture generic or service-specific behaviors. In addition, the developer can define rules based on metrics collected from one or more VNFs in one or more NFVIs to receive runtime notifications. Furthermore, authors in [40] proposed an approach for comprehensive and detailed monitoring of 5G mobile networks characterized by software using an IoT-based system. The corresponding monitoring framework is designed to collect any type of data either in text or in numerical form in a cloud database. Thus, by using this knowledge, SDN controllers make the decisions on network reconfiguration according to current conditions. A great work was published also in [41]. In this article, a Software Defined Monitoring (SDM) was proposed, while it highlights how SDMs can be used to solve the current limitations in legacy monitoring systems. The proposed approach is able to monitor both virtualized and physical network environments in an economical and efficient way. Initially, the authors' proposed SDM architecture was used only to monitor 5G backhaul network. Last but not least, an automatic monitoring management for 5G mobile networks was proposed in [42]. In particular, a 5G-oriented architecture was proposed to integrate SDN and NFV technologies to monitor and control the entire service life cycle taking into account network control plane information. This architecture automatically manages network resources to orchestrate network monitoring services, which are developed in their solution as VNF monitoring. VNF monitoring is described by the collection of information from different and diverse sources, such as network (i.e. physical or virtual) infrastructure data, network management services and user- to- network communication

2.3. NFV-Catalogues & Recommendations

The conception of the NFV Catalogues for the coherent storage of the exchanged entities of the 5G infrastructure originates from several applications into the new era of virtualization. The first attempt of delivering a distributed storage component with functional business and service layers for the VNFs/NS operators was from the [43]. The main view of this approach was to provide a digital marketplace that collects VNFs / NS to operate on commodity cloud infrastructures. On top of that, the continuous and real-time network information of the available VNFs/NS is exchanged between the several layers of the NFV Catalogue based on a set of RESTFul APIs for a functional service. Though, the first concrete approach was presented from in the framework of the T-NOVA project [44]. The authors presented the storage of the machine-readable descriptions of the VNFs/NSs and was covered solely from the NFV Catalogues as an integrated component. In parallel, the NFV Catalogues were fully aligned the functional components of the T-NOVA infrastructure, responsible for the charge of registering all business relationships and exposure of the related information for the billing component. Although, this approach presents a distributed storage approach with the several QoS/QoE metrics being disperse in the ecosystem, introducing latency in the diverse functionalities. The next milestone in the evolution of the NFV Catalogues was set one more time by the SONATA project. This approach was predominantly based on the efficient storage of the generalized package formats of the NFV landscape and their functional information. The followed approach was strictly correlated with the specification of the ETSI MANO for the diverse employed NFV Catalogues in the infrastructure [45]. What was beyond the specification comprised the introduction of the engagement of the information on the instantiated VNFs/NSs in accordance to the SONATA Service Platform policy updates.

In parallel with the vast development of the Web, the advent of the rapid growth of the available information was obvious to its users. Recommender Systems (RSs) pioneered the web with the aim of incorporating social information and at the same time delivering meaningful suggestions to the end user. While the research field of RSs has been skyrocketed in diverse domains, there is a gradual, yet slow, interest of the application of the RSs in the 5G ecosystems, through their pinpointing in network management applications [46-48]. Through the introduction of the virtualized era, telecom networks generate massive amounts of monitoring data consisting of observations on network faults, configuration, accounting, performance, and security. E- stream also included a predictive, automated network management recommendation with surprising results due to the ever- increasing complexity of the networks, correlated with particular business level constraints [49]. Through the exploitation of the profound streams of data and the efficient application of techniques in dimension reduction, E-stream was based on recommending

actions with four different aligned factors, namely the context, audience, existing responses, and validation. The main factor of these recommendations was the applicable trading-policies for actuating the recommendation module and propelling the necessary actions to get the best price for each VNF be carried out. Yet, in the 5G telecom systems, there is a paramount absence of RSs in utilizing implicitly the plethora of the QoS/QoE metrics of the multiple diverse components, such as monitoring systems, policy and SLA modules, etc. Thus, RSs techniques and methods comprise a subset of tools that need to be examined thoroughly.

2.4. The challenges of 5G & SLAs

An SLA is a contract between a service provider and its end users that documents what services the service provider can provide and establishes the performance standards that the service provider is required to meet [50]. Since each component potentially impacts the overall behavior of the SDN, any high-level target specified for the service (e.g. performance, availability, security) potentially impacts all low- level components.

SLAs establish customer expectations regarding the service provider's performance and quality. In recent years, SLAs have set expectations for the performance of a service provider, set down penalties for missing targets and, in certain cases, rewards for exceeding those targets [21]. With SDN, network devices (routers and switches) can be managed using OpenFlow [51] and a separate set of application programming interfaces (APIs) can handle virtual network overlays. Once the network is virtualized, the SDN controller can set up network devices as fast as it can deploy new VMs. For instance, customers can take a VM image, deploy it to hardware, spin it, apply it via OpenStack Compute and set up the network around it in a short time via OpenStack Neutron (i.e. Orchestrated SDN form) [52]. In this scenario, SLAs can cover the time and cost of deploying new compute resources, as well as their related network resources [53].

Finally, the SLA lifecycle is an important part of the provision of services, in particular in SDN and 5 G networks. The SLAs lifecycle in the 5 G domain is managed by the accompanying 5 G service platforms and is slightly different from the traditional ones as described in [54]. SLA management is a dynamic process comprising four key stages: a) Architecture, b) Engagement, c) Operations and d) Termination as presented in Figure 1. The overall lifecycle is 5G-enabled, due to the fact that is fully aligned with 5G principles and is running in parallel with Network Service Lifecycle [55].

The first phase starts with the selection of a NS and the requirements definition by the developer. Typically, the Operator is the one responsible to examine those, take into consideration important business needs and implement SLA Templates, as initial offers to the NS customers. During the engagement phase, the selection of different NS results from business aspects, which are the basis for different QoS constraints, which can also be defined as requirements of the agreement.

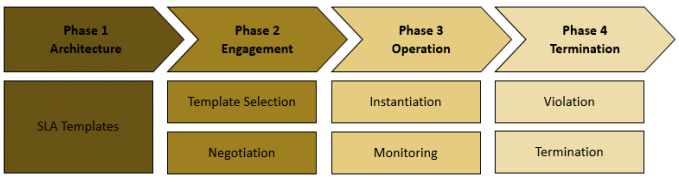


Figure 1 : SLAs Lifecycle in 5G.

The preference of an operator / network service provider depends on the desired network service (NS), its characteristics, budget constraints and so on. QoS expectations also drive end users to negotiate with their operators / service providers precise QoS levels. An SLA is created to describe the agreed QoS parameters after a successful negotiation process. After the successful NS selection and the corresponding SLA negotiation process, the operation phase takes place. This phase comprises the actual deployment of the NS, the population of the respective service with running data, the establishment of communication channels and additional operational activities. Moreover, the operation phase monitors the agreement with real- time data, for the purpose of avoiding or managing unexpected violations. Finally, termination phase deals with the end of the relationship between operator/service-provider and NS customer, including the end of the legal relationship. In general, the latter will continue for a few years after termination in accordance with mandatory laws and legislation. This last phase includes the evaluation of alternatives, settlement and termination commitments, export of data, customer care and diligence, and deletion of data. All the above should be considered either the Network Service was terminated, or the SLA was violated. More details for the aforementioned processes and how they are managed in the proposed architecture will be discussed in Section 3 of this article.

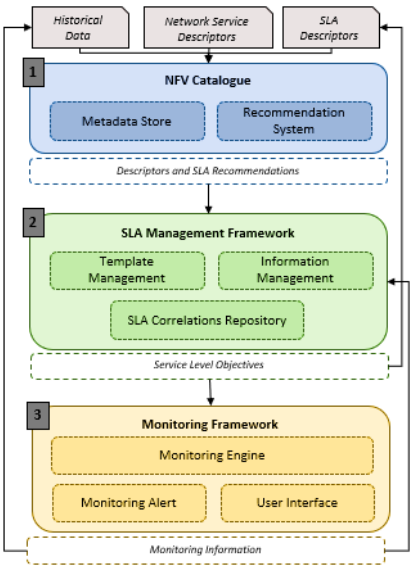


Figure 2: Overall framework architecture.

3. Proposed SLA-Oriented Framework

Considering the challenges presented in Section 2, the developed framework allows to manage the whole lifecycle of SLAs, from the template formulation to the violation detection. Figure 2 depicts the overall architecture of the developed framework. Taking into consideration the inter-connection between a distributed set of users and resources, the implementation of an "SLA-Oriented Framework" aims at governing this interaction. Due to the fact that all running service affects the performance of the service platform, any business parameter (i.e. high-level requirements) included in the SLA, would be linked to recourse demands – encapsulated in the respective policies (i.e. low-level requirements). Initially, a set of NSs, SLAs and historical

monitoring data among with their metadata, are stored into the "NFV-Catalogue". This preliminary information is used from the "Recommendation System", to provide feedback toward the "SLA Management Framework" for an optimized SLA Template formulation. As soon as this information is available, the SLA Framework should be considered in two phases: a) the SLA Template Management and b) the SLA Information Management. During the first phase, an optimized SLA Template is prepared, by mapping the high-level expressed by the customer into low-level resource attributes needed by the service provider. On the other hand, during the second phase, where the NS instantiation takes place, our proposed framework oversees the obtain monitoring data from the "Monitoring Framework", which lies on the bottom of the proposed architecture. The implemented "Monitoring Framework" provides support for the QoS management, while it provides adaptable monitoring feedback based on the infrastructure needs. The following sub-sections explain in detail each component.

4. NFV Catalogue

4.1. Metadata Store

As an initial stage, the "NFV-Catalogue", which is depicted in Figure 2, is positioned to address storage and management necessities of diverse stakeholders' (i.e. NS developers, NS providers, customers etc.). The main view of the "NFV-Catalogue", is to provide a repository for persistent storing of the developed VNFs/NS and their corresponding SLAs, attaching them with additional metadata, which are exploited to leverage its functionalities and interfaces for storing, searching and retrieving. Moreover, additional information, like metric significance outcomes and information related to policies and QoS need to be also stored as metadata of the corresponding VNFs. Thus, the "NFV-Catalogue" is deemed to be a multi-faceted data storage, addressing various stakeholder needs while also forming the main and centralized data storage of the 5G ecosystem. The functionality of the "NFV-Catalogue" is predominantly based on the metadata for NS Descriptors and SLA Descriptors. Prior to the attachment of the metadata, through a RESTful API of the "NFV-Catalogue", the inspection of the validity of the document structure is a critical and necessary step. Since the documents are specified in machine-readable formats, the review of the format contributes to the eliminations of flaws in the "NFV-Catalogue". Moreover, the attachment of metadata provides the ability of defining uniquely the individual stored machine-readable objects inside the data storage as depicted in Figure 3. Moreover, "NFV-Catalogue"

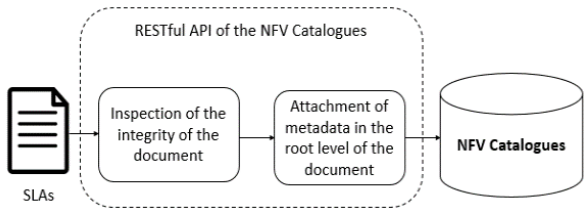


Figure 3: SLAs and metadata storage.

is aligned with the principle of the persistent storage by extending this type of information with valuable fields for successful data integration, accuracy in the format of the document, confirmed time of creation, etc. In this way, it enables the development of enhanced operations for Creating, Retrieving, Updating and Deleting (CRUD) SLA Descriptors inside the "NFV-Catalogue", while reassures the correct data format of the stored documents (e.g. SLA Templates).

Going beyond the conventional data storage, the presented "NFV-Catalogue" provide intelligent functionalities in a 5G environment. Since the types of information vary, one of the necessities that is satisfied by the Catalogues is the full-text search capabilities in structure-agnostic documents. Since the schema of the diverse documents (i.e. NS descriptors, SLA descriptors) is variable, the "NFV-Catalogue" provides searching capabilities without the necessity of indexes. Thus, it provides seamless retrieval abilities in deep-hierarchical machine-readable document structures. Furthermore, besides from the plain NoSQL document store for the diverse descriptors, "NFV-Catalogue" provides a scalable file system for hosting the artifact files, required for the instantiation lifecycle of the VNFs/NSs.

4.2. Recommendation System

After the successful storage of the NS Descriptors among with their SLAs, we consider a "Recommendation System", which will optimize the SLA Templates creation, by providing important knowledge to the Operator/NS-Provider. Thus, through the plethora of the stored SLAs and the historic data of each end-user, it is undisputedly valuable to tap into them and provide optimum combinations of the available QoS parameters through recommendations. The basic principle of these recommendations is that paramount dependencies presented between the user-to-item activity. The aim of the "Recommendation System" is not to provide only specific and optimum combinations of QoS but also to allow users to profit from them. Thus, the recommendations comprise samples from the relevant actions that were followed in several signed SLAs from similar end-users. In the proposed framework, user-based Collaborative Filtering (CF) was used in order to detect similar users and promote recommendations in these terms [56]. The user is deemed that explicitly rated a combination of SLAs, along with the included QoS parameters. Moreover, what is of paramount importance is the metric of computing the correlations between the end users in the "Recommendation System". More specifically, the Pearson correlation metric was the most appropriate for the best trade-off in the equation of quality-number of predictions [57]. Although, the aforementioned filtering suffers from two severe issues, the Data Sparsity and the Cold Start [58]. The former, comprises the phenomenon that end users provide a small amount of ratings, contributing to memory complexity and inevitability in training the "Recommendation System". The latter refers to the difficulty in bootstrapping the "Recommendation System" for newly-introduced users and items. Despite this fact, the CF, in the proposed "SLA-Oriented Framework", eliminating these challenges by introducing "trust relationships" of end-users. The main idea is the provision of the trust metric for each individual user-to-user and, on top of that, contributes to the predictions of the conventional CF, as depicted in Figure 4. The trust metric relevance of the users is denoted from the received implicit rating of each VNF/NS. With the instantiation/upload of a VNF/NS, the "NFV-Catalogue" automatically receives implicit rating of the respective entity. Thus, the selections can provide the relevance of the diverse users of the platform.

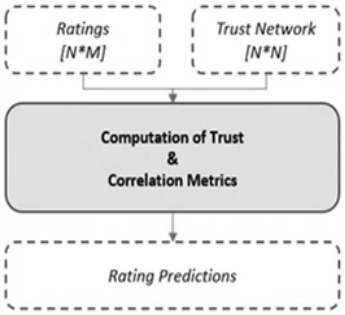


Figure 4: Proposition of trust metrics

5. SLA Management Framework

In the second stage of the proposed "SLA-Oriented Framework", the SLA Templates are generated, as an initial offer to the end-users. After the successful NS instantiation, along with the corresponding SLA, the signed Agreement starts to be monitored so it can fulfill the signed SLOs to the end-user. The current stage is splitted into two sub-phases, a) SLA Template Management, which takes place prior the NS deployment, and b) the Information Management, which takes place during the NS deployment. The internal architecture and workflow inside the "SLA Management Framework" it is depicted in Figure 5.

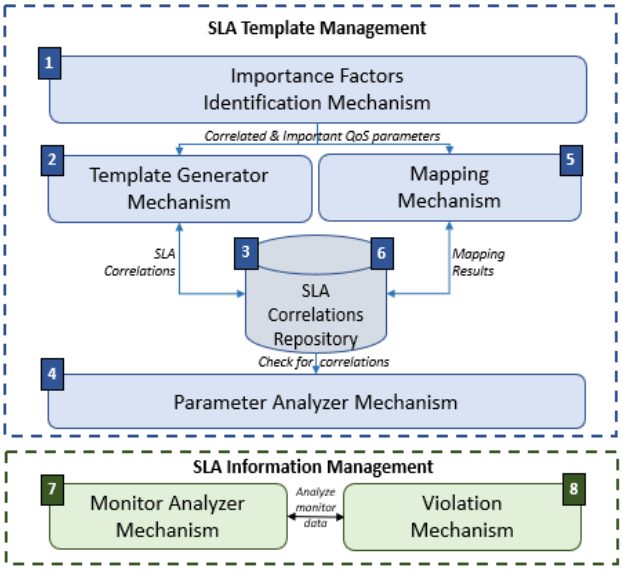


Figure 5: SLA management framework internal architecture

5.1. SLA Correlation Repository

Given the big amount of data generated and analyzed into the SLA Management Framework, an internal repository has been introduced, namely "SLA Correlations Repository", in order to store and manage all the necessary correlations, between end-users, network services, templates, agreements, violations, as well as recourse mapping results. In particular, the correlations between the high-level and the low-level requirements are stored for future analysis. In addition, it keeps track of all the correlations between the generated templates and the linked network services. At the same time agreements information are also located in the repository, along with the end-user's authentication details, as well as the violations records.

5.2. SLA Template Management

The SLA Template Management is the first phase of the SLA Management component while it is in charge of receiving the desired business guarantees (i.e. high-level QoS parameters) from various parties (e.g. NS provider, customer), and formulate an initially SLA Template. It is also responsible for mapping the high-level business guarantees to low level resource attributes, so they are able to be included in the Template, and afterwards monitored through the instantiated Agreement. The SLA Template Management consists of four mechanisms which are going to be further described in the following sub-sections.

Importance Factors Identification Mechanism

In the case of NSs, a challenge arises given that many different entities are setting their requirements for the overall service. Those entities may have specific preferences for resource attributes and potentially additional parameters that can be monitored (e.g. number of sessions) and thus be included into an SLA. To address the aforementioned challenge, we would need to develop a mechanism which would analyze monitoring data and performance information in order to identify dependencies between a VNF's metrics but also realize how these dependencies affect the overall performance of a specific NS. The latter would be reflected to the so called "importance" factors, while it could give feedback to the Operator, about what it is of crucial importance and need to be included in the SLA Template. This is in fact an on-line learning process that updates and dynamically evolves. To this end, the corresponding mechanism, estimates and defines the importance of various QoS parameters through correlation analysis of a) historical SLA parameters, b) predefined policies, c) historical monitoring data and d) recommended QoS metrics. The latter it is depicted in Figure 6.

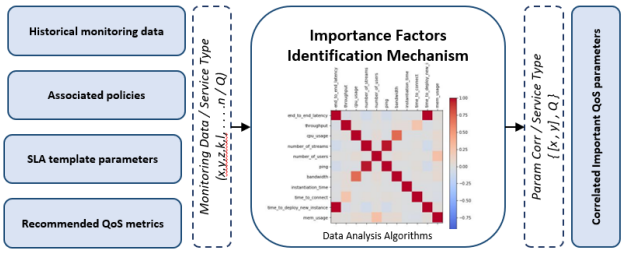


Figure 6: Importance factors identification mechanism

As one understands, the color of the correlogram is white when there is no correlation between the two variables (the correlation value is equal to 0 or close to 0). When the color is deep black it means that our mechanism has calculated a sufficiently large positive correlation, while the deepest grey color indicates there is a large negative correlation.

It should be mentioned that policies are considered out of the proposed framework's scope, and that's why it is assumed that are able to be provided by an external Policy Management Framework. In order to correlate the QoS parameters and define the important ones, analysis libraries were used to analyze the performance measurements of the chained VNFs in a NS. Starting an offline learning process with the gathered data, the mechanism calculates and stores the dependencies in an internal NoSQL database. Then, providing REST APIs as well as a GUI, the end-

user can request and get those dependencies of his/her developed NS, and as a result getting an inside knowledge of the NS's performance behavior. Thus, the "Importance Factors Identification Mechanism" can produce essential weight factors and classify parameter dependencies, to suggest and include relevant QoS parameters in the SLA templates. [59-61].

Template Generator Mechanism

As soon as the important QoS parameters are recognized, the "Template Generator Mechanism" takes action, which initially produces the SLA Templates requested by the service provider, and then it is responsible to establish the final Agreement. The "Template Generator Mechanism" as shown in Figure 7, can acquire a set of policies for a clearly defined NS and also historical data of the service provider through the "Monitoring Framework" (i.e. NS performance data, preferences of resource parameters) [62].

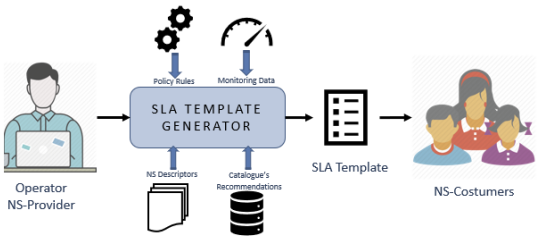


Figure 7: SLA template generator mechanism

In addition, the "Template Generator Mechanism" accesses the "NFV-Catalogue" to retrieve also the NS Descriptor for the corresponding NS, as well as QoS parameters recommendations from the Catalogue's RS. Finally, an also important input to the "Template Generator Mechanism", is the weight factors of the aforementioned recommended QoS parameters, obtained by the "Importance Factors Identification Mechanism". After gathering all the above-mentioned input, the "Template Generator Mechanism", triggers the analysis while it correlates the input in a way that it can formulate an SLA Template with some initial guarantees.

Mapping Mechanism

As soon as the initial Template is formulated, there is essential need to decompose the business guarantees to recourse attributes that can be monitored accordingly by the "Monitoring Framework". In other words, The SLA "Mapping Mechanism" (MM) is the component responsible to translate the high-level requirements described by the end-user into low-level metrics required by the service provider, and vice-versa. More specifically, the MM obtains a set of policy rules from an external Policy Management Framework, a set of low-level and high-level requirements described from the service provider but also from the customer. As a result, the produced output of the MM (i.e. output layer of the ANN) are explicit SLA business metrics, as depicted in Figure 8.

The MM is based on unsupervised learning, using an Artificial Neural Network (ANN) [63]. ANNs can be used to solve this translation problem by mapping service-specific SLOs to resource attributes directly. As they embody a black box approach, ANNs are ideal to be used in an environment where information is not easily transmitted from one entity to another. In addition, ANNs need no knowledge of the inner structure of the NSs [64]. However, it should be noted that they need a representative execution dataset, in order to detect complex, linear or non-linear dependencies.

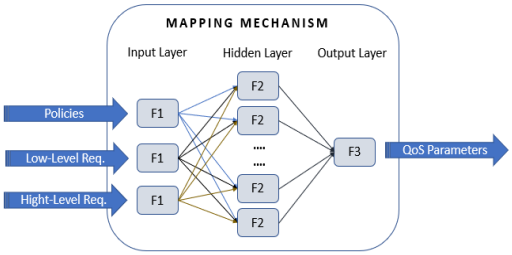


Figure 8: ANN based mapping mechanism

Parameter Analyzer Mechanism

The "SLA Parameter Analyzer" it is capable of deciding whether or not the MM should occur. The decision is made after the component searches into the "SLA Correlations Repository" and check whether the input parameters already correlate with the mapping results calculated and stored.

5.3. SLA Information Management

Once the selected NS has been successfully instantiated, the SLA Information Management should take action by continuous monitoring the service instance. This process aims to optimize the SLA formulation, manage the guaranteed terms in respect with the infrastructure conditions, while also check for any SLA violations. SLA Information Management consists of two mechanisms which are going to be presented explicitly in the next sub-sections. The first is called "SLA Monitor Analyzer Mechanism" while the second "SLA Violation Mechanism".

Monitor Analyzer Mechanism

Starting with the "SLA Monitor Analyzer Mechanism" the component acquires both historical monitoring data for the deployed NS and biases for resource parameters. Sequentially, it should decide whether there is any existed contrast between the mapping results and the runtime monitoring data. Specifically, the "SLA Monitor Analyzer Mechanism", examines the QoS parameters from the "SLA Correlations Repository", with the gathered monitoring data, while it afterwards measures the delta between their values. The MM will obtain the monitoring feedback as an additional dataset and re-train the ANN, in case the delta is greater than "0".

Violation Mechanism

As a final step, in the second stage of the proposed "SLA-Oriented Framework", we consider the "SLA Violation Mechanism". The corresponding mechanism is responsible to ensure that the newly deployed NS would not violate the corresponding SLA upon instantiation, while it successfully fulfills the signed business needs. Nevertheless, the mechanism is also responsible to identify any violations occurred and take the necessary actions (e.g. scale in, scale out). Specifically, upon receipt of the measured metrics through the Monitoring Framework, the "SLA Violation Mechanism" starts an ongoing

process of re-adoption. In particular, the mechanism considers the mapping results stored in the "SLA Correlations Repository" and compare them with the real-time monitoring information. In anticipation of future SLA violation threats, the mechanism readjusts the low-level recourse parameters described for the SLA and push them back to the "Monitoring Framework". In this way, both the SLA is fulfilled, while at the same time the infrastructure does not waste recourses, but only in times of real need. Although, in case a violation is not prevented in time, an alert is sent from the "Monitoring Alert Mechanism". Upon receival, the "SLA Violation Mechanism" calculates the overall value of the specific metric and takes decision whether the SLA is violated or not. In case of an SLA violation, the customer is informed by an e-mail, SMS, or even a live push notification.

6. Monitoring Framework

In the third, and final, stage of the proposed "SLA-Oriented Framework", lies the "Monitoring Framework", where, its internal architecture is depicted in Figure 9. The proposed monitoring framework has adopted the SONATA Monitoring Framework [65], and then adapted accordingly in order to support the whole NS lifecycle in respect of the associated SLAs. The corresponding monitoring framework is consisted of: a) the "Monitoring Engine", which collects monitoring data provided by the NSs based on the signed SLA, b) the "Alert Manager", which is responsible to produce alert messages when a violation of a SLA rule is occurred, and c) a "User Interface", used for visualization of the collected monitoring data, while also visualization of each individual rule specified by the "SLA Management Framework".

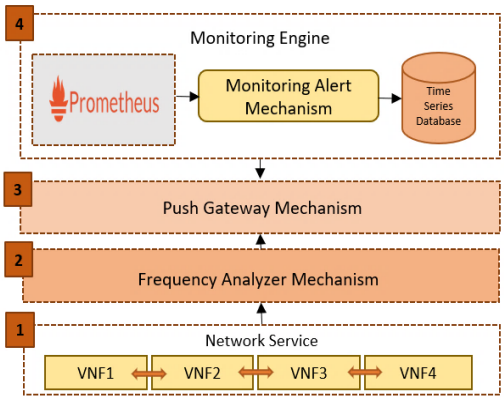


Figure 9: Monitoring Framework: Internal Architecture

6.1. Monitoring Engine

To begin with, a key mechanism of the third stage is the "Monitoring Engine", on the bottom of the proposed "SLA-Oriented Framework". We should point out that, the automation of monitoring an SLA, is a difficult task that demands precise specifications and an adaptable mechanism that collects the right measures and models. At the same time, evaluation of an SLA should occur in specific time frames or when some remarkable events happen. In 5G/SDN ecosystem, where chained VNFs in form of NSs are implemented and deployed on top of a service platform, it becomes essential to create a "Monitoring Engine" which is able to manage a variety of specifications and monitor accordingly the recourses of the virtualized infrastructure. While an SLA is already attached and instantiated through the NS, it is assumed that the desired guarantee terms have been granted to the customer. On the instantiation phase of the service, monitoring rules for the specific NS instance are automatically generated and pushed to the Monitoring Framework, through the "SLA Violation Mechanism". The "Monitoring Engine", which is based on a stable version of Prometheus Monitoring System [66] will undertake the collection of monitoring data from the running services. Prometheus scraps metrics for short- lived jobs either directly or through an intermediate push gateway. It stores all scraped samples locally and executes rules on these data to either add new time series from available data or generate alerts. The benefit of using Prometheus as "Monitoring Engine" is the fact that it was designed for reliability, and the ability to allow quick problems diagnosis.

Push Gateway Mechanism

As it is previously mentioned, Prometheus, can scrap monitoring data exported from the running service instances through the push gateway. To be more specific, the Push Gateway, is a subcomponent of Prometheus, acting like an intermediary service, allowing to forward the monitored data from the "Monitoring Engine" towards the "Monitoring Alert Mechanism", and thus publish them to the external components (i.e. SLA Management Framework, NFV-Catalogue). In our case, the usefulness of the "Push Gateway Mechanism" arrives during the NS scaling up [67]. Scaling up, means that a new VNF is about to start, relieving the service instance when it is actual needed. In this case, the new VNF identifies the "Push Gateway" in terms of authentication and push the monitoring data towards a recognized and reliable host. As a result of the approach, is the fact that the "Monitoring Engine" does not need to know and identify the NS instances, but vice-versa.

Frequency Analyzer Mechanism

After the previous discussion around pushing monitoring data through the "Push Gateway", a challenge arises, in terms of how often and which data are promoted to the engine. Having this in mind, an important parameter of define the above-mentioned challenge is the time interval used to evaluate the resource metrics and guaranteed SLOs (e.g. every two seconds or every two minutes). Although, too frequent pushes may affect negatively the overall system performance, whereas too infrequent pushes may cause heavy SLA violations, due to lack of monitoring metrics towards the "SLA Management Framework" [68]. To this end, an enhancement to Prometheus Monitoring Framework is introduced, namely the "Frequency Analyzer Mechanism", which is based to an adaptive monitoring algorithm. Thus, the "Frequency Analyzer Mechanism" acts as a middle agent between an active connection of the NS instance, the "Push Gateway" and the "Monitoring Alert Mechanism" [17]. Its purpose is to provide highly accurate information about the network's health, while at the same time avoid the production of unnecessary traffic in the network. It aims at adapting during runtime the monitoring time intervals in order to ensure that the data collected and transmitted to the SLA Management Framework, are fruitful and not all raw data. Moreover, it should be noted that the algorithm achieves significant reduction in resource consumption and also reduces the number of SLA violations, due to the pro-active nature of the mechanism.

Time Series Database

Finally, a key component of the Monitoring Framework, is also an internal database. A "Time Series Database" is used for storing and identifying the monitored information, by a metric name and a set of key-value pairs. Following the approach of a time series database, the advantage of having operators for calculating useful information of the monitoring data, is given to the "Monitoring Engine".

6.2. Monitoring Alert Mechanism

The previously discussed monitoring outcomes are going to be published to the external components (i.e. SLA Management Framework, NFV-Catalogue) though the "Monitoring Alert Mechanism". During our research we realized that a message queue system (MQ) was the most appropriate solution for the intercommunication of the aforementioned components. Therefore, RabbitMQ was integrated to the Monitoring Framework, as the message broker for asynchronous messaging [69]. Through the implemented "Monitoring Alert Mechanism", SLA monitoring rules and SLA violations are produced as alerts. Thus, the message (i.e. alert) is pushed to the "Monitoring Alert Mechanism", and all the RabbitMQ consumers are receiving the message for further actions. It should be noted that, one of the consumers is the "SLA Violation Mechanism". This mechanism, acts as an intermediate component between the "Monitoring Engine" and the end-user of the NS instance.

6.3. User Interface

For the visualization of the gathered monitoring data, Grafana is used as an open platform for visualize and beatify monitoring data analytics [70]. Grafana, features an advanced chart query editor that lets the user to quickly browse the metric space, add features, change operating parameters, and more.

7. Evaluation

In order to evaluate the performance of the proposed framework in terms of efficiency and ease of use, our approach was included in the innovative 5G infrastructure environment of the 5GTANGO Service Platform [71]. 5GTANGO project is an EU funded Innovation Action, that enables the flexible programmability of 5G networks with a modular Service Platform so it can bridge the gap between business needs and network operational management systems [72]. The 5GTANGO Service Platform offers the service and functional orchestration features, along with all the supplementary and supporting tools required, like the proposed "NFV-Catalogue", "SLA Management Framework" as well as the "Monitoring Framework".

7.1. Emulation Environment

During the evaluation of the proposed framework a challenge arised, as many of the mechanisms (i.e. Recommendation System, Importance Weight Factors Mechanism, Mapping Mechanism) need apriory behavior knowledge, in order to be able to deal with unknown VNFs/NSs and train their models properly. Moreover, this becomes even more necessary in the emerging DevOps environments, where new versions of NSs are directly deployed in production (i.e. working environment), and therefore no up-to-date monitoring data is available for the updated services. To deal with this challenge, we adopted the OSM supported VIM emulator [77], in order to run on top of it the 5GTANGO VNF/NS benchmarking framework, to automatically execute performance benchmarks of NFV network services and functions [78, 79]. The benchmarking tool automatically gathers performance information about a service, prior to its deployment without requiring dedicated testbeds, resulting to an offline profiling of the service, and the collection of benchmarking data, so they can used as a starting point of the service modeling.

7.2. Working Environment

As it is previously mentioned, the proposed "SLA-Oriented Framework" is implemented inside the 5GTANGOs' SP where the installation guide can be found in [80]. For evaluation purposes, we used the NCSR Demokritos' testbed in order to setup the SP and thus our proposed framework along it. NCSR Demokritos' testbed is the main node of the 5GTANGO infrastructure in Athens, providing the following infrastructure components: a) WAN network, b) Access network, c) datacenter (computing resources for NFVI realization), and d) end user devices and services. To be more specific, as Network Function Virtualization Infrastructure (NFVI) Queens OpenStack multi node deployment with provider networks configuration is used [82]. Also, service chaining is taken care by a Service Function Chaining (SFC) agent that interfaces with the Service Platform in order to fix the chaining between the Network Service components. For the Wide Area Network (WAN) part, the networking is managed by a WAN Infrastructure Manager (WIM) implemented by a Virtual Tenant Network (VTN) running on top of OpenDayLight (version Oxygene) [83]. Finally, the current study used a processing environment which consists of the following elements:

- One Dell R210 used as the Fuel jump host
  - 1xIntel(R) Xeon(R) CPU X3430 @ 2.40GHz
  - 4GB RAM
  - 1TB HDD
- One Dell T5500 functioning as the Controller
  - 2xIntel(R) Xeon(R) CPU X5550@2.67GHz
  - 16GB RAM
  - 1.8 TB HDD
- Three Dell R610 utilized as Compute Nodes
  - 2xIntel(R) Xeon(R) CPU E5620@2.40GHz
  - 64GB RAM
  - 1.8 TB HDD
- One Dell R310 used as NFVI-PoP
  - 1xIntel(R) Xeon(R) CPU X3450@2.67GHz
  - 16GB RAM
  - 465GB HDD

More details about the testbed's topology, hardware/software availability and network recourses can be found in [73]. It should be also noted that with regards to the source code availability, it is

currently partially in open-source format, since not all functionalities have finalized yet.

7.3. Experimental Results

In this case study, different stakeholders take place to the overall workflow, in order to provide the necessary inputs. In this study, we consider a) Thomas – the Service Developer, b) Sally – the Commercial Offer Designer (COD), responsible for defining SLAs supporting the whole business, c) Bob – the network engineer of the Service Provider in charge of defining run-time policies, d) Robert – the engineer, in charge of providing monitoring data and e) End-user Customers, such as Brian. The deployment of a network service instance in a service provider's infrastructure comes with the definition of some requirements. This will ensure the performance estimation and the QoS requested by the customer. Quality of service is introduced in the SLA with the definition of SLOs, along with policies for managing the infrastructure accordingly, and thus enforces the respective SLA.

In order to perform a complete testing and evaluation of our approach, an elastic proxy network service is used as depicted in Figure 10, which is consisted of two chained VNFs: a) a HAProxy VNF, configured as a load balancer and b) a Squid VNF configured as a proxy server.

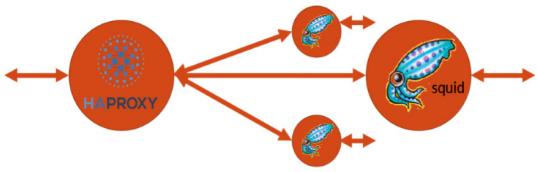


Figure 10: Elastic proxy network service

The end-users use the ingress interface of the HAProxy as proxy IP, and sequentially the HAProxy forwards the incoming requests to one of the Squids in its backend pool. To begin with, the NS-Descriptor, as developed by Thomas, the service developer, is onboarded to the proposed "NFV-Catalogue". The onboarding is made through a Rest API (i.e. Create operation), as mentioned in Section 3.1. Then, the Sally, the Commercial Offer Designer, is responsible to define the customers' facing characteristics of the service, in particular the SLAs. The SLA Generator requires four (4) parameters for a successful generation of the SLA template. The most critical one is the selection of the NS (i.e. the haproxy-squid NS) that is going to be correlated with the newly created SLA template. At the same time an SLA name among with a valid future expiration date and at least one SLO are needed. Once this information is gathered, the SLA Generator mechanism triggers the "Importance Weight Factors Mechanism" and the "Recommendation System". At this moment, the first one will feed the "Template Generator Mechanism" with feedback on *"relevant"* QoS parameters that need to be included in the template, and at the same time, Catalogue's "Recommendation System" will also provide recommendations based on previous created templates and relevant costumers' preferences. It is important to point out that, computing this kind of recommendations and similarities on behalf of Sally, enhance the generation process, by giving the chance to the Commercial Offer Designer to include in the template the most appropriate QoS parameters, and minimize the negotiation between Brian (i.e. the customer) and the service provider. Tables 1 presents high level parameters included in previous customers' SLAs, along with their calculated and monitored values, while Table 2 depicts the similarity between Sally and previous costumers of the haproxy-squid service. It should be mentioned that the values presented in Table 1, are real business requirements, as gathered from Communication Pilot use case of the 5GTANGO project [76].

Table 1: Previous customers low-level metrics

	Customer A	Customer B	Customer C
Availability	0,99.99 %	0.95 %	0.9 %
Jitter	10 ms	25 ms	50 ms
Packet Loss	0.1 %	0.05 %	0.5 %

Table 2: Sally's similarity with previous customers

	Sally	Customer A	Customer B	Customer C
Sally	1	82 %	50 %	30 %

Based on the above results, the "Recommendation System" of the "NFV-Catalogue" and the "Importance Weight Factors Mechanism", Sally is recommended to include the following guarantees to the SLA Template, as presented in Table 3.

Table 3: QoS recommendations for Sally

	Availability	Jitter	Packet Loss
Sally	99 %	15 ms	0.1

Afterwards, the generation of the SLA template, as an initial offer to the NS-costumers is triggered. The outcome is an SLA Descriptor that will be onboarded to the NFV-Catalogue, and it is based on WS-Agreement specification [74-75]. The main SLA Templates building blocks of the reference model include the root element, the SLA template and the service elements, as depicted in Figure 11.

Afterwards, Brian, the end-user customer, browses through the available NSs, select the haproxy-squid NS and instantiate it, in order to be deployed in the SP. During the instantiation process the customer triggers automatically the one-shot negotiation process through the "Mapping Mechanism", by selecting the previously generated SLA Template for the specified NS. Based on the business requirements of the customer, the SLA can be accepted, or a new negotiation process can be initiated. Once the SLA Manager has collected all the relevant datasets via the "Parameter Analyzer", checks if there is already a combination in the "SLA Correlations Repository", between the latter and the already existing mapping results, in order to decide whether the process of the "Mapping Mechanism" should be triggered or not. In case there is not a correlation yet, the operator's low-level requirements, the costumer's high-level business needs and the policies, are mapped in order to produce the actual QoS parameters that can finally be included in the SLA. The objective is to forecast the performance and the quality that is required, to be agreed and signed in the final SLA. Alternatively, if there is an already a

combination between the input dataset (i.e. requirements obtained from the operator and the costumer) and the stored mapping results, the SLA Manager bypasses the mapping process and dynamically creates the final Agreement. In order to investigate this mapping, we needed to predict the performance of the Network Service on top of the infrastructure.


```
{
  "name": "silver-template-example",
  "vendor": "UPRC",
  "version": "2.0",
  "author": "Evgenia Kapassa, Marios Touloupou",
  "description": "This is a Gold SLA Template for Haproxy-Squid Service",
  "sla_template": {
    "template_name": "Gold",
    "offer_date": "2019-02-04T11:35:10Z",
    "expiration date": "2020-02-04T11:35:10Z",
    "provider_name": "Telefonica",
    "template_initiator": "Evgenia Kapassa",
    "service": {
      "ns_uuid": "0e69ccfd-d9ba-4439-99b8-cd4f2a059457",
      "ns_name": "ns-squid-haproxy",
      "ns_vendor": "eu.5gtango",
      "ns_version": "0.2",
      "guaranteeTerms": [
        {
          "guaranteeID": "g1",
          "guarantee_name": "Availability",
          "guarantee_threshold": "99",
          "guarantee_operator": "greater",
          "guarantee_unit": "%",
          "guarantee_period": "Daily",
          "guarantee_definition": "",
          "guarantee_service_level": "50sec/24h",
          "target_slo": [
            {
              "target_kpi": "Downtime",
              "target_value": "50s",
              "target_operator": "less",
              "target_duration": "10s",
              "target_period": "24h",
              "target_service_level": "Downtime less 50s"
            },
            {
              "target_kpi": "Jitter",
              "target_value": "15 ms",
              "target_operator": "less",
              "target_duration": "10s",
              "target_period": "",
              "target_service_level": "Jitter less than 15 ms"
            },
            {
              "target_kpi": "Packet Loss",
              "target_value": "0.1",
              "target_operator": "%",
              "target_duration": "",
              "target_period": "",
              "target_service_level": "0.1% Packet Loss of the total packets sent"
            }
          ]
        }
      ]
    }
  }
}
```


Figure 11: SLA descriptor example

The output is categorized between simple mapping results and complex ones. A simple mapping result maps "end-to-end", from low-level to high-level. For instance, mapping the low-level metric "downtime" to high level SLA parameter "availability". Complex mapping results include predefined formulations to calculate specific SLA parameters using low level resource metrics. Table 4 presents an example of a complex mapping result.

Table 4: Complex mapping result example

Low-Level Metric	SLA Parameter	Mapping Formulation
downtime, uptime	Availability (A)	$A = 1 - \frac{downtime}{uptime}$

In order to investigate the performance of the "Mapping Mechanism", the emulation environment provided a data set consisted of 360 data points. Of these, 50% was used in order to train the network model, for the haproxy-squid NS, as depicted in Figure 12, resulting into 180 data points. For validation purposes another 20% was used during training, meaning 72 data points. Finally, the overall network capability measured against the remaining 30% of the data set, (on which the model was not trained), simulating a real test situation.

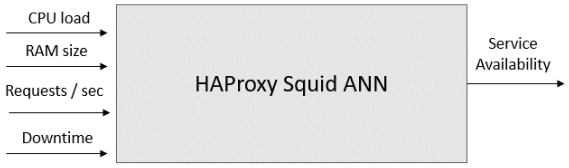


Figure 12: ANN model for HAProxy-Squid Network Service

Then, the remaining 30% was used to check the system's reliability in accurately predicting the NS's QoS levels in the Deployment Environment. In order to check the absolute differences between the ANN's prediction and the actual monitoring observation, the Mean Absolute Error (MAE) was used, as shown in (1), where 'n' is the number of data points, y_j represents the observed values and $\hat{y}_j$ the predicted values. The MAE result is depicted in Table 5.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \quad (1)$$

Table 5: Complex mapping result example

ANN Model	Neuron per Layer	Mean Absolute Error
HAProxy-Squid NS Availability	4 – 3 - 1	2.75 %

The aforementioned ANNs are feed-forward back-propagation networks, trained with the Levenberg-Marquardt algorithm [81]. The criterion for performance was the Mean Square Error (MSE) in the training set, while it was trained for 100 periods, for a training time of 1 minute. Putting all this together, we have the general formula for calculating the MSE in (2).

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \quad (2)$$

It should also be stated that, if the customer wishes to change the recourse parameters that the HAProxy-Squid service can handle, the 5GTANGO platform, which incorporates the service presented in this paper, offers the capability of re-triggering the "Mapping Mechanism" and retrain the model.

Next, the instantiation of the NS takes place. To this end, the deployment aims to enable execution of the NS according to the QoS requirements, while at the same time appropriate monitoring, allows the measurement of QoS parameters at both service and infrastructure levels targeting events of resource provisioning estimation and decision making. For this reason, the Monitoring Framework access the Point of Presence (PoP) that the NS is deployed into and gathers monitoring information for the haproxy-backend-downtime, jitter and packet loss, in order to measure its availability. At this point, the "Frequency Analyzer Mechanism" takes place, by adjusting the monitoring time intervals during runtime to ensure that the data collected and transmitted to the SLA management framework are meaningful. At first, data are collected and compared with linear increase of time intervals, until they rich an initial time threshold. In our case the collected monitored values

of the HAProxy-Squid service downtime were below the certain threshold, indicating that the network has changed towards a better state. Therefore, after the service had pushed the monitored data, the "Frequency Analyzer Mechanism" multiply the time the change occurred, with β, in order to increase the data transmission interval. As a consequence, the new timeout is higher, and the probes will collect data with a linear increase over a longer time period, without wasting recourses.

As depicted in Figure 13, the X axis shows a linear increase in data collection time from the samples. The latter begins in the first second and increases linearly until it reaches the fifth second. while a change in the metric value occurs in the 5th sec. The "Frequency Analyzer Mechanism" commands the probes to send their data to the "Push Gateway", and at the same time increases the timeout by the current time* β (i.e. current time is the time when the significant change occurred).

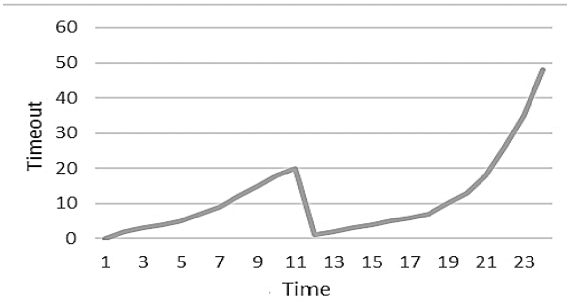


Figure 13: Time intervals adjustments

In our scenario, the equation returned the value of 18 seconds, meaning that the data collection will continue after a linear increase in time until the new timeout which set to 18th second is reached. Moreover, the monitoring process of the proposed approach was tested firstly using the standard Prometheus framework, and then by enhancing it with the aforementioned "Frequency Analyzer Mechanism". Figure 13 depicts the difference regarding the network workload (i.e. throughput in terms of data requests towards the push gateway per second). Finally, when the network service has completed its lifecycle, Brian, is responsible for terminating it. The termination process, as well as all the aforementioned procedures, are taking place in a user-friendly way, through a unified Portal.

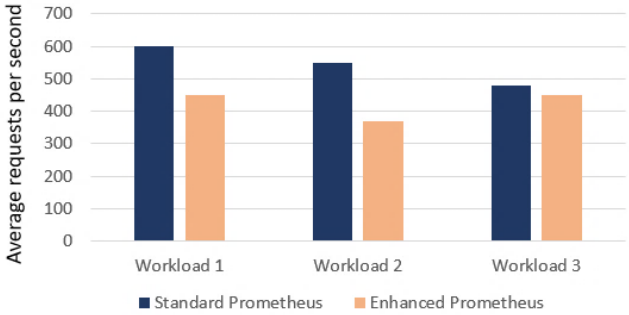


Figure 13: Network workload with and without using the "Frequency Analyzer Mechanism"

8. Conclusions

Based on the aforementioned evaluation, we presented a SLA Management Framework that is used to map high- level business parameters to low- level attributes of resources. This framework is integrated in the 5GTANGO Service Platform for the autonomous management of SLAs. We considered a generic approach that is based on ANNs in order to efficiently be used as a mediator for the network provider and the end-user. We considered an ANN based approach that can be used as a mediator between the end-user and the provider. Furthermore, we have introduced a mechanism to determine the importance of various QoS parameters so that the "relevant" ones be included in the SLAs for better QoS assurance. The proposed monitoring framework goes a step forward from the traditional implementation, by preventing any unnecessary traffic to the network, and also by providing real-time and high accurate information for better QoS assurance. Last but not least, the authors have presented a beyond a plain data storage to an enriched information – driven repository, the so-called "NFV-Catalogue". The aggregation of the stored information allows the mechanism to apply recommender system techniques, build on QoS predictions and SLA recommendation systems. Regarding the three major components of the proposed architecture (i.e. the NFV-Catalogue, SLA Management Framework and Monitoring Framework), we conclude to the following, based on the captured experimental results.

All things considered, provisioning of resources in a virtualized 5G infrastructure is a challenging task, that still needs a lot of investigation. Therefore, we plan to extend the framework in order to enable Quality of Experience (QoE) enforcement. This kind of enforcement could be done by adopting the infrastructure recourses accordingly during runtime, considering parameters based on Catalogue's recommendations as well as monitoring feedback. Moreover, we tend to enhance the SLA violations management, by providing violations prediction models, in order to prevent day zero violations. Additionally, the currently proposed framework is able to monitor and manage business guarantees in a single` 5G environment. Therefore, we envision to manage a 5G network with multiple domains, enabling higher level of integration, and at the same time adapt the proposed architecture from a wide range of verticals, enabling higher level of abstraction.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been partially supported by the 5GTANGO project, funded by the European Commission under Grant number H2020ICT-2016-2 761493 through the Horizon 2020 and 5G-PPP programs (http://5gtango.eu).

References

[1] F. Hu, Opportunities in 5G Networks: A Research and Development Perspective, CRC Press, 2016.
[2] A. Osseiran et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," in IEEE Communications Magazine, 52(5), 26-35, 2014. doi: 10.1109/MCOM.2014.6815890
[3] H. Su, X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings Over Cognitive Radio Wireless Networks," in IEEE Journal on Selected Areas in Communications, 26(1), 118-129, 2008. doi: 10.1109/JSAC.2008.080111
[4] D. Wu, R. Negi, "Effective capacity: a wireless link model for support of quality of service," in IEEE Transactions on Wireless Communications, 2003. doi: 10.1109/TWC.2003.814353

[5] C. Liang, F. R. Yu and X. Zhang, "Information-centric network function virtualization over 5g mobile wireless networks," in IEEE Network, 29 (3), 68-74, 2015. doi: 10.1109/MNET.2015.7113228

[6] Q. Zhu, X. Zhang, "Game-theory based power and spectrum virtualization for maximizing spectrum efficiency over mobile cloud-computing wireless networks," in 49th Annual Conference on Information Sciences and Systems (CISS), 2015. doi: 10.1109/CISS.2015.7086818

[7] Samsung Developers, "5G Requirements", Online: https://developer.samsung.com/tech-insights/5G/5g-requirements

[8] Huawei Technologies Co. Ltd, "5G Network Architecture: A High-Level Perspective", Online: https://www.huawei.com/minisite/hwmbbf16/insights/5G-Nework-Architecture-Whitepaper-en.pdf

[9] E. Casalicchio, V. Cardellini, G. Interino, M. Palmirani, "Research challenges in legal-rule and QoS-aware cloud service brokerage", in Future Generation Computer Systems, 78(1), 211-223, 2018. doi: https://doi.org/10.1016/j.future.2016.11.025.

[10] Anuta Networks, "Top 6 Challenges for Service Assurance in NFV", Online: https://www.anutanetworks.com/top-6-challenges-for-service-assurance-in-nfv/

[11] A. Tabebordbar, A. Beheshti, "Adaptive Rule Monitoring System" in 1st International Workshop on Software Engineering for Cognitive Services, 2018. doi: 978-1-4503-5740-1

[12] L.B. López, J. M. Vidal, L.G. Villalba, "An Approach to Data Analysis in 5G Networks", in Entropy, 19(2), 74, 2017. doi: 10.3390/e19020074

[13] Centina, "Service Assurance Critical to SDN/NFV Success", Online: http://www.centinasystems.com/service-assurance-critical-sdnnfv-success/

[14] E. Kapassa, M. Touloupou, A. Mavrogiorgou, D. Kyriazis, "5G & SLAs: Automated proposition and management of agreements towards QoS enforcement" in 21st Conference on Innovation in Clouds, Internet and Networks and Workshops, 2018. doi: 10.1109/ICIN.2018.8401587

[15] E. Kapassa, M. Touloupou and D. Kyriazis, "SLAs in 5G: A Complete Framework Facilitating VNF- and NS- Tailored SLAs Management," in 32nd International Conference on Advanced Information Networking and Applications Workshops, 2018. doi: 10.1109/WAINA.2018.00130

[16] S. Benkner and G. Engelbrecht, "A Generic QoS Infrastructure for Grid Web Services," in Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services, 2006. doi: 10.1109/AICT-ICIW.2006.16

[17] M. Touloupou, E. Kapassa, A. Kiourtis and D. Kyriazis, "Cheapo: An algorithm for runtime adaption of time intervals applied in 5G networks," in Fifth International Conference on Software Defined Systems. doi: 10.1109/SDS.2018.8370420

[18] N. Sfondrini, G. Motta, L. You, "Service level agreement (SLA) in Public Cloud environments: A Survey on the current enterprises adoption," in 5th International Conference on Information Science and Technology, 2015. doi: 10.1109/ICIST.2015.7288964

[19] Xi Zhang, Jia Tang, Hsiao-Hwa Chen, Song Ci and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," in IEEE Communications Magazine, 44(1), 100-106, 2006. doi: 10.1109/MCOM.2006.1580939

[20] L. Zhu, X. Liu, "Technical Target Setting in QFD for Web Service Systems Using an Artificial Neural Network," in IEEE Transactions on Services Computing, 3(4),338-352, 2010. doi: 10.1109/TSC.2010.45

[21] V. C. Emeakaroha, I. Brandic, M. Maurer, S. Dustdar, "Low level Metrics to High level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," in International Conference on High Performance Computing & Simulation, 2010. doi: 10.1109/HPCS.2010.5547150

[22] T. Cucinotta et al., "Virtualised e-Learning with real-time guarantees on the IRMOS platform," 2010 IEEE International Conference on Service-Oriented Computing and Applications, 2010. doi: 10.1109/SOCA.2010.5707166

[23] G. Kousiouris, D. Kyriazis, S. Gogouvitis, A. Menychtas, K. Konstanteli, T. Varvarigou, "Translation of application-level terms to resource-level attributes across the Cloud stack layers," in IEEE Symposium on Computers and Communications, 2011. doi: 10.1109/ISCC.2011.5984009

[24] K. Samdanis, X. Costa-Perez, V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," in IEEE Communications Magazine, 54(7),32-39, 2016. doi: 10.1109/MCOM.2016.7514161

[25] X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in IEEE Communications Magazine, 55(5), 94-100, 2017. doi: 10.1109/MCOM.2017.1600951

[26] V. Tikhvinskiy, G. Bochechka, "Prospects and QoS Requirements in 5G Networks", in Journal of Telecommunications and Information Technology, 1, 23-26, 2015. doi:

[27] ETSI, "ETSI TR 102 889-2 V1.1.1, Technical Report", Online: https://www.etsi.org/deliver/etsi_tr/102800_102899/10288902/01.01.01_60/tr_10288902v010101p.pdf

[28] G.C. Madueño, C. Stefanović, P. Popovski, "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling" in IEEE Wireless Communications Letters, 3(4), 429-432, 2014. doi: 10.1109/LWC.2014.2326674

[29] J. Tang X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation over Wireless Links," in IEEE Transactions on Wireless Communications, 6(8), 3058-3068, 2007. doi: 10.1109/TWC.2007.051075

[30] A. Asadi, V. Mancuso, "A Survey on Opportunistic Scheduling in Wireless Communications," in IEEE Communications Surveys & Tutorials, 15(4), 1671-1688, 2013. doi: 10.1109/SURV.2013.011413.00082

[31] T. Guo, R. Arnott, "Active LTE RAN Sharing with Partial Resource Reservation," in IEEE 78th Vehicular Technology Conference), 2013. doi: 10.1109/VTCFall.2013.6692075

[32] K. Hammad, A. Moubayed, S. L. Primak and A. Shami, "QoS-Aware Energy and Jitter-Efficient Downlink Predictive Scheduler for Heterogeneous Traffic LTE Networks," in IEEE Transactions on Mobile Computing, 17(6), 1411-1428, 2018. doi: 10.1109/TMC.2017.2771353

[33] J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí, "On the configuration of radio resource management in a sliced RAN," in IEEE/IFIP Network Operations and Management Symposium, 2018. doi: 10.1109/NOMS.2018.8406280

[34] I. da Silva et al., "Impact of network slicing on 5G Radio Access Networks," in European Conference on Networks and Communications, 2016. doi: 10.1109/EuCNC.2016.7561023

[35] A. Tootoonchian, M. Ghobadi, Y. Ganjali, OpenTM: Traffic Matrix Estimator for OpenFlow Networks, Springer, 2010

[36] C. Yu, Lumezanu, Y. Zhang, V. Singh, G. Jiang, H.V. Madhyastha, FlowSense: Monitoring Network Utilization with Zero Measurement Cost, Springer, 2013.

[37] M. Yu, J. Yu, L. Rui, M. Rui, Software Defined Traffic Measurement with OpenSketch, USENIX Association, 2013.

[38] P. Trakadas et al., "Scalable monitoring for multiple virtualized infrastructures for 5G services" in The International Symposium on Advances in Software Defined Networking and Network Functions Virtualization, 2018. doi: http://hdl.handle.net/1854/LU-8569066

[39] SONATA Project Consortium, "SONATA NFV: Agile Service Development and Orchestration in 5G Virtualized Networks", Online: http://www.sonata-nfv.eu/

[40] T. Maksymyuk, S. Dumych, M. Brych, D. Satria, M. Jo, "An IoT Based Monitoring Framework for Software Defined 5G Mobile Networks" in Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, 2017. doi: 10.1145/3022227.3022331

[41] M. Liyanage et al., "Software Defined Monitoring (SDM) for 5G mobile backhaul networks," in IEEE International Symposium on Local and Metropolitan Area Networks, 2017. doi: 10.1109/LANMAN.2017.7972144

[42] A. H. Celdrán, M.G. Pérez, F. J. García Clemente, G. M. Pérez, "Automatic monitoring management for 5G mobile networks" in Procedia Computer Science, 110, 328-335, 2017. https://doi.org/10.1016/j.procs.2017.06.102.

[43] N. Nikaein. et al., "Network store: Exploring slicing in future 5g networks" in Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture, 2015. doi: 10.1145/2795381.2795390

[44] T-NOVA Project Consortium, "T-NOVA: Network Functions As-a-Service Over Virtualized Infrastructures", Online: http://www.t-nova.eu/

[45] M. Ersue, "ETSI NFV Management and Orchestration - An Overview", Online: https://www.ietf.org/proceedings/88/slides/slides-88-opsawg-6.pdf

[46] R.E. Núñez-Valdéz, et al., "Implicit feedback techniques on recommender systems applied to electronic books" in Computers in Human Behavior, 28(4), 1186-1193, 2012. https://doi.org/10.1016/j.chb.2012.02.001.

[47] K. Oku, R. Kotera, K. Sumiya, Geographical Recommender System Based on Interaction Between Map Operation and Category Selection, ACM, 2010.

[48] J. Serrano-Guerrero, E. Herrera-Viedma, J.A. Olivas, A. Cerezo, F.P. Romero, "A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0" in Information Sciences, 181(9), 1503-1516, 2011. doi: https://doi.org/10.1016/j.ins.2011.01.012.

[49] F. Zaman, G. Hogan, S. V. Der Meer, J. Keeney, S. Robitzsch, G. Muntean, "A recommender system architecture for predictive telecom network management," in IEEE Communications Magazine, 53(1), 286-293, 2015. doi: 10.1109/MCOM.2015.7010547

[50] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, C. Pu, "An Analysis of Performance Interference Effects in Virtual Environments," in IEEE International Symposium on Performance Analysis of Systems & Software, 2007. doi: 10.1109/ISPASS.2007.363750

[51] ONF, "Software-Defined Networking (SDN) Definition", Online: https://www.opennetworking.org/sdn-definition/

[52] OpenStack Project, "OpenStack Networking ("Neutron")", Online: https://wiki.openstack.org/wiki/Neutron

[53] B. Hoff, “IT service-level agreements and SDN: Assuring virtualization performance”, Online: https://searchnetworking.techtarget.com/tip/IT-service-level-agreements-and-SDN-Assuring-virtualization-performance
[54] SLA-Ready, “Cloud SLA lifecycle”, Online: http://www.sla-ready.eu/cloud-sla-lifecycle
[55] S. Van Rossem et al., "A network service development kit supporting the end-to-end lifecycle of NFV-based telecom services," in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2017. doi: 10.1109/NFV-SDN.2017.8169859
[56] C. Saluja, “Collaborative Filtering based Recommendation Systems exemplified”, Online: https://towardsdatascience.com/collaborative-filtering-based-recommendation-systems-exemplified-ecbffe1c20b1
[57] J. Jin, S. Zhang, L. Li, T. Zou, "A Novel System Decomposition Method Based on Pearson Correlation and Graph Theory" in IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS), 2018. doi: 10.1109/DDCLS.2018.8515967
[58] X. He, L. Liao, H. Zhang, L. Nie, T. Chua, “Neural Collaborative Filtering”, in Proceedings of the 26th International Conference on World Wide Web, 2017. doi: 10.1145/3038912.3052569
[59] Z. u. Rehman, F. K. Hussain, O. K. Hussain, "Towards Multi-criteria Cloud Service Selection," in Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011. doi: 10.1109/IMIS.2011.99
[60] Z. ur Rehman, O. K. Hussain, S. Parvin, F. K. Hussain, "A Framework for User Feedback Based Cloud Service Monitoring," in Sixth International Conference on Complex, Intelligent, and Software Intensive Systems, 2012. doi: 10.1109/CISIS.2012.157
[61] T. Halabi, M. Bellaiche, "Evaluation and selection of Cloud security services based on Multi-Criteria Analysis MCA," in International Conference on Computing, Networking and Communications, 2017. doi: 10.1109/ICCNC.2017.7876216
[62] B. K. Tripathy, A. G. Sethy, P. Bera, M. A. Rahman, "A Novel Secure and Efficient Policy Management Framework for Software Defined Network," in IEEE 40th Annual Computer Software and Applications Conference, 2016. doi: 10.1109/COMPSAC.2016.31
[63] S. Hussain, R. Atallah, A. Kamsin, J. Hazarika, “Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study” in Silhavy R. (eds) Cybernetics and Algorithms in Intelligent Systems, Springer, 2019. doi: https://doi.org/10.1007/978-3-319-91192-2_21
[64] G. Kousiouris, D. Kyriazis, S. Gogouvitis, A. Menychtas, K. Konstanteli, T. Varvarigou, "Translation of application-level terms to resource-level attributes across the Cloud stack layers," in IEEE Symposium on Computers and Communications, 2011. doi: 10.1109/ISCC.2011.5984009
[65] SONATA Project Consortium, “D4.3 Service Platform First Operational Release and Documentation”, Online: http://www.sonata-nfv.eu/sites/default/files/sonata/public/content-files/deliverables/SONATA%20D4.3%20Service%20platform%20operational%20release%20and%20documentation.pdf
[66] rometheus Authors, “Prometheus: From metrics to insight”, Online: https://prometheus.io/
[67] S. Dutta, T. Taleb, A. Ksentini, "QoE-aware elasticity support in cloud-native 5G systems," in IEEE International Conference on Communications (ICC), 2016. doi: 10.1109/ICC.2016.7511377
[68] C. Vincent et al., “Towards autonomic detection of SLA violations in Cloud infrastructures”, in Future Generation Computer Systems, 28(7), 1017 – 1029, 2012. doi: https://doi.org/10.1016/j.future.2011.08.018.
[69] Pivotal Software, “RabbitMQ”, Online: https://www.rabbitmq.com/
[70] Grafana Labs, “Grafana: The open platform for beautiful analytics and monitoring”, Online: https://grafana.com/
[71] C. Parada et al., "5Gtango: A Beyond-Mano Service Platform," in European Conference on Networks and Communications, 2018. doi: 10.1109/EuCNC.2018.8443232
[72] 5GTANGO Project Consortium, “5GTANGO: 5G Development and Validation Platform for global Industry-specific Network Services and Apps”, Online: https://5gtango.eu/
[73] 5GTANGO Consortium, “D6.1 Infrastructures, Continuous integration approach”, 2017, Online: https://5gtango.eu/project-outcomes/deliverables/38-d6-1.html
[74] A. Andrieux et al., “Web Services Agreement Specification (WS Agreement)”, Grid Resource Allocation Agreement Protocol (GRAAP) WG, 2011, Online: https://www.ogf.org/documents/GFD.107.pdf
[75] R. Kabert, G. Katsaros, T. Wang, "A RESTful implementation of the WS-agreement specification", in Proceedings of the Second International Workshop on RESTful Design, 2011. doi: 10.1145/1967428.1967444
[76] SONATA Project Consortium, “Real Time Communications”, Online: https://5gtango.eu/index.php/about-5g-tango/47-real-time-communications
[77] M. Peuster, H. Karl, S. v. Rossem, “MeDICINE: Rapid Prototyping of Production-Ready Network Services in Multi-PoP Environments” in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2016. doi: 10.1109/NFV-SDN.2016.7919490
[78] M. Peuster, H. Karl, “Profile Your Chains, Not Functions: Automated Network Service Profiling in DevOps Environments” in IEEE Conference on Network Function Virtualization and Software Defined Networks, 2017. Doi: 10.1109/NFV-SDN.2017.8169826
[79] M. Peuster, H. Karl, “Understand Your Chains: Towards Performance Profile-based Network Service Management” in Fifth IEEE European Workshop on Software Defined Networks, 2016. doi: 10.1109/EWSDN.2016.9
[80] 5GTANGO Project Consortium, “5GTNAGO Service Platform Installation Guide” Online: https://sonata-nfv.github.io/component_installation
[81] G. Kousiouris et al., “Distributed Interactive Real-time Multimedia Applications: A Sampling and Analysis Framework” in 1st International Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems, 2010. doi: https://eprints.soton.ac.uk/id/eprint/272323
[82] OpenStack Project, “OpenStack Queens Expands Support for GPUs, Containers to Meet Edge, HA, AI Workload Demands”, Online: https://www.openstack.org/software/queens/
[83] OpenDaylight Project, “OpenDaylight”, Online: https://www.opendaylight.org

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 339-353 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Recent Advances in Intelligent-Based Structural Health Monitoring of Civil Structures

Satyam Paul[*,1], Raheleh Jafari[2]

[1]School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Nuevo Leon, 64849, Mexico

[2]Centre for Artificial Intelligence Research (CAIR), University of Agder, Grimstad, 4879, Norway

ARTICLE INFO

Article history:
Received: 16 August, 2018
Accepted: 10 October, 2018
Online: 18 October, 2018

Keywords:
Structural health monitoring
Damage detection
Civil structures
Intelligent techniques

ABSTRACT

This survey paper deals with the structural health monitoring systems on the basis of methodologies involving intelligent techniques. The intelligent techniques are the most popular tools for damage identification in terms of high accuracy, reliable nature and the involvement of low cost. In this critical survey, a thorough analysis of various intelligent techniques is carried out considering the cases involved in civil structures. The importance and utilization of various intelligent tools to be mention as the concept of fuzzy logic, the technique of genetic algorithm, the methodology of neural network techniques, as well as the approaches of hybrid methods for the monitoring of the structural health of civil structures are illustrated in a sequential manner.

1. Introduction and Preliminary Concepts

The main intention of structural health monitoring methodology is to the expose of the area of damage and then to evaluate its location and severity thus predicting the life. Visual inspection and experimental measurement techniques are an important aspect of damage detection methodology. More and more research on the field of global detection methods in the identification of damages in the structure during and aftermath of seismic events has resulted in the investigation of methodology used in mechanical, nuclear, and aerospace engineering [1, 2]. The civil engineering structures like multistory buildings, towers and bridges tend to damage gradually with time elapse or due to sudden environmental disasters. So structural health monitoring is an important area of research that verifies the safety of civil structures from various issues. The stability and reliability of civil structure becomes a matter of concern after a strong earthquake and so more and more studies are carried on structural damage detection using different nondestructive computation methods. The vibration techniques are another popular method to identify the location of the damages with accuracy and thus considered to be global technique. The vibration methods associated with structural health monitoring detects the changes in the modal properties of the structure.

The constraints associated with these techniques are: a) the influence of factors producing same to same effects on the evaluation parameters which are not easy segregate, b) civil engineering structures has superfluity that give rise to low sensitivity of the method in case of localized damaged, c) the uncertainties in the model, d) the noises in the measurements, e) in case of training data, uncertainties may incorporated in the data due to adverse conditions and d) in case of artificial neural networks, there is completeness of the conceived damage scenarios and training data sets [3, 6]. The method of damage detection of civil structures using non-destructive techniques (NDE) is an area of interest among the researchers. Its applications extents from aviation industries, manufacturing techniques, space shuttles, architectural industry, devices related to power plants, etc. The damage detection in structures cab be subdivided into local and global techniques. The local methodology are mainly applied for tiny and generalized structures for example pressurized vessels. The limitation of damage identification method involving local concepts is that it fails to detect the damages in case of large and complicated structures. In case of onsite structural damage detection, engineers can apply local detection method only to investigate some special component of the structures. The analysis of entire structure including some large and complicated structure is possible through global vibration-based identification of damages in structure [7, 8]. The variation in the parameters of structural modal is considered to be an early signal of the occurrence of damage in the structures [9]. The validation of reliability in function of numerous crucial technical structures by tracking online and detection of damage in continuous mode is possible using vibration based methodology. As a result of which vibration based method is growing area of interest among the

*Satyam Paul, Monterrey, Mexico, satyam.controlsystems@gmail.com

researchers. The main concept of vibration based damage detection is that the changes associated with the physical properties like mass, damping and stiffness will result in a change in the modal properties like natural frequency, modal damping and mode shapes. The past three decades have witnessed the contribution of the researchers in all three domains of time, frequency and modal domains. The effort of the researchers in this field will continue as existing methods cannot rectify all the damages that is incorporated in the structures [10]. In [11], the author has broadly distinguished between various methods on the basis of the level of identification attempted as shown in Table 1:

Table 1: Level of identification

Level	Various Methodology
1	To determine that the damage exists in the structure
2	To determine the damage geometric location
3	To estimate the level of extent of damage
4	To forecast the remaining life of the structure

Vibration-based structural damage detection can be broadly classified into traditional and modern type which is still under development and innovation. The description and comparison of traditional method and modern methods have been vividly described in forthcoming sections. The modern-type methods of structural health monitoring can be classified as wavelet analysis, fuzzy logic concept, neural network techniques, genetic algorithm methodology. In this paper, the intelligent technique related to structural monitoring techniques are discussed and analyzed. The intelligent methodologies can sub divided as neural network techniques, Fuzzy logic theory, concept of genetic algorithm and hybrid intelligent techniques. The researchers are taking lot of interest in the optimization methodology involving genetic algorithms [12] and neural networks [13] in order to investigate localized damage variation as a function of modal data**.** In the area of research associated to civil structures to the optimization of structure[14] as well as scheme related to the identification of structure [15] there has vast use of genetic algorithm for handling optimization problems. The detection of structural damages using this methodology is growing vastly. The synergistic combination of algorithms associated with Eigen sensitivity and genetic algorithm for optimizing the discrete variables of damage area as well as their extent utilizing natural frequencies is illustrated by [16]. The methodology of genetic algorithm in the detection of structural damages was presented by [17]. They utilized optimization fitness function techniques associated to displacements which is static in nature. Good linear mapping capability, inverse problems solving ability and high degree of robustness is the criteria of artificial neural network that makes it an important methodology in the field of structural health monitoring and detection. Several investigators utilized artificial neural network for the detection of localize as well as to compute the damage involved. The popular type of neural network termed as multi-layer perception (MLP) is utilized in health monitoring of structures as well as well in the situations involving identification [18, 20]. In the area of fault diagnosis and structural damage detection, there is a wide popularity of fuzzy logic for its embedded abilities in extracting precise and reliable data from imprecise as well as uncertainty features [21, 22]. Fuzzy logic has been employed in wide areas of fault diagnosis and structural damage detection and is a growing area of interest. The developed fuzzification technology associated with fuzzy concept is counted for uncertain factors as well as the methodology of training approaches is uncounted in numerable phase involving noisy determinants in comparison to the neural approaches for accounting of uncertainty [23]. In [24], the authors demonstrated a concept to evaluate structural damage utilizing vibration data as well as fuzzy clustering. A novel methodology was demonstrated by [25] which is a combination of fuzzy logic, neural network and wavelet transformation for the detection of damages in structures. Researchers are taking widespread interests in the field of structural health monitoring that uses combined intelligent methodologies. In [26], the authors utilized a fuzzy-neuro hybrid technique to monitor the health condition of bridges. The decision system of fuzzy logic utilizes fuzzy clustering methodology in order to determine the damages in the bridge. A neural network forecasting technology makes use of back propagation algorithm to predict the correct amount of damage. An innovative filter design on the basis of combined neural and genetic algorithm concept for the improvisation health monitoring technique was proposed by [27].

In the next section, focused is vested on the traditional techniques associated to structural health monitoring. This section is followed by a modern methodology involved in health monitoring od structures which is followed by a brief overview of wireless sensors and then detailed analysis on intelligent techniques are illustrated. Finally conclusion is provided to give an overview of the paper.

2. Methods based on traditional techniques

The conventional methodology for the detection of damages in structure utilizes the mechanical approaches related to the structural attributes. The methodology involved in this type of techniques are transfer function computation or experimental modal evaluation. This techniques are not widely used in online detection of structures because the experimental data requires instruments which are diversified or involves manual operations. Therefore this technology has been not taken up with greater interest in recent studies. The main advantage of natural frequency analysis technique is that it is easily measurable in case of structures and does not depend on the measured position. Generally the preciseness of computation utilizing natural frequency methods is superior when compared with the mode shape analysis or modal damping. In [28], the authors proposed an innovative technique for extracting the information on the size and location of structural damage by utilizing the concept of natural frequency. In the current method, the location of the crack is obtained using Armon's Rank-ordering method and selecting the first four natural frequencies. In [29], the researchers proposed an innovative technique for the detection and computation of the damaged layout in structural devices utilizing a limited natural frequency. An algorithm with the properties of localized damages criteria is generated for tracking damages in terms of natural frequency variance as well as quantifying damage algorithms in order to predict the dimension of the crack. The method solely verifies the presence of large damage, but it is a matter of difficulty

to extract the damage location due to the fact that structural damage in different location may result in same frequency alteration. The changes in natural frequency will result in detection of damages in structure but in order to access the location of structural damage, information on vibration modes is utter necessary. In [30], the authors handled the constraints associated with the extraction of damages utilizing outcome based computation of vibration in the grip of varying environmental conditions. Using the computation techniques, two variety of features are extracted categorize under Eigen properties of the structure with help of automatized stochastic identification of the subdivided space methodology as well as with the help of peak indications calculated based on Fourier transform associated with the filters related to the modal. The method of modal analysis for the research on the damages on a wooden wall was investigated by [31]. They use the technique of damage-sensitive parameters computation for the investigation**.** The damaged zone is detected by inspecting visually the various period of damaged states related to the deformation in mode dimensions. The analysis of modal residue as well as variations in stiffness is an requirement for superior illustration of the damaged zone. The structural vibration mode is lower when compared with the natural frequency associated with the preciseness of computation although vibration mode generate more information related to damages. The damage detection methodology based on flexibility matrix change depends on the mode shape as well as reciprocal natural frequency and also the generalized criteria is satisfied by the structural vibration mode. Hence, in flexibility matrix the high frequency element will decrease in a rapid manner with the increasing of natural frequency. In [32], the author proposed a methodology for diagnosing damage in structure on the basis of change in dynamically computed flexibility and structural stiffness. The detection of damage is not only the primary objective, the detection of location is also very much essential. The implementation of the identification methodology utilizing covariance-driven subspace gives an idea of structural modal parameters. The main idea is to extract the dimensions of the matrix that are flexible and is related to the computed DOF. The concept of pseudo inversion will yield stiffness matrix. The damage localization is extracted with combined analysis of changes in two calculated matrices considering a reference state until the damaged state. In most cases, when some damage is visible in a structure, it is convenient to extract more information from stiffness matrix than the mass matrix. The one of the properties of stiffness is that it changes in astonishing manner when major damage occurs in a structure and therefore the change of stiffness matrix can be used to detect damages. But this method fails to do justice in case the damage is very small. Non-inclusion of mathematics model, vibration mode measurement and experiential knowledge in frequency response function methods makes it an advantageous model. So this model can be used for monitoring online. In [33], the researcher proposed and developed new monitoring techniques considering higher order spectra for the frequency response functions in order to monitor involved nonlinearities in structure as well as non-Gaussian types of signal in case of damage as well as the computation of related to the phase-coupling of harmonics signal associated with the structures. The proposed methodology is in general traditional functions associated to the frequency response for the investigation of spectrum for actual value transformations in higher order. A damage identification technique relied on artificial neural network in combination with frequency response functions was investigated by [34]. This technique can effectively analyze nonlinear damages suitable to some level of excitation**.** The important concept of their work is to illustrate a practical methodology associated to vibration dependent structural health monitoring. This innovative technique reduces the dimension of the initial frequency response function values and converts it into latest damage indices and engages technique of artificial neural network for the exposure of different level of nonlinearity with the help of damage patterns recognition. The main disadvantages of this methodology is that the quantity as well as the zone of measured point have an impact on the accuracy of structural damage detection. A newly developed concept for the identification of structural damages based on the dual methodology of instant identification of damping coefficient and wavelet transform was presented by [35]. The combination of experimental results and numerical simulations justifies that in structural systems with incorporated damages results in crucial variations in damping. Thus, the parameters linked to the structural damping can be used as sensitivity to damages associated to system properties. In [36], the authors proposed a technique where sensitivity-based finite element (FE) model updating methodology is implemented for damage detection. The optimization algorithm is utilized for the minimization of the objective function and to initiate the damage detection techniques. In [37], the authors developed a novel damage indicator termed to be as mode shape area index on the basis of mode shape changes. This developed methodology found out to be as the most sensitive damage detection approach. A statistical technique with combined uncertain frequency as well as mode shape data for the detection of damages in structure is suggested by [38]. The suggested technique is implemented to a laboratory tested steel cantilever beam and frame structure. The experimental results validate that all the damages are detected suitably with high probabilities of damage existence. The utilizing of frequency response functions for the detection of damages in structure was proposed by [39]. The suggested technique iteratively maximizes the elements related to the damages by not including the elements which were validated as undamaged from the sets of good elements. The mentioned approaches was applied to numerical simulations using a structural setup having 2D frame, also the efficiency was established. These conventional techniques have some constraints to be mention as [40]:

- Due to its dependency on experimental results to be mention as modal shape computation as well as the calculation involving damping but it incurs computation cost. This creates a negative impact on online damage detection techniques.
- In terms of damage location identification, this methodology is not popular because of its complexity to develop a recognized methodology considering several structures due to its severe dependency on the acquirement of the individual structures which are required for the detection.
- The conventional techniques fail to unearth the minute damages associated with the structures.

3. Detection techniques based on modern methodologies

Modernized pattern of vibration reliable structural damage detection which is also known as intelligent damage extraction

technique is a renowned methodologies for the detection of damage utilizing online computed structural vibration responses. The methodology utilize combined technology of current signal processing concepts as well as artificial intelligence analyze mechanisms. The methods fall under this category is wavelet analysis, neural network, genetic algorithm, etc. Wavelet analysis is an upcoming research filed mainly in the area of signal processing. It has been implemented with success in many applications such as transient signal analysis, image analysis, structural health monitoring system etc. The existence of damage in structure can be identified using the spectrum graph obtained using wavelet transform. A wavelet based concept for the detection of structural damages was proposed by [41]. In their work, they illustrated the analysis of vibration signals characteristics associated with wavelet transformations. The concept is implemented to simulated data extracted from an complex structural model under the effect of harmonic excitation. The model is made up of numerous breakable springs, of few may undergo irreversible damage when the value of response exceeds the value of threshold or the number of cycles of motion is stored beyond their fatigue span.

In [42], the authors demonstrated an innovative concept on the basis of fuzzy concept involving wavelets phenomena for the identification of damages in structure. The methodology is the combination of Wavelet Packet Transform which is used for feature extraction and the capabilities of fuzzy sets to model vagueness as well as uncertainty. The operation are carried out in two steps mainly the training of pattern and the monitoring of health. The general difference illustrating the importance of each technique is given in Table 2. The elaborate explanation about the intelligent methodologies associated with structural health monitoring has been depicted vastly in this survey paper.

4. Importance of wireless sensors for the health monitoring of structures

The technology of wireless sensing is an important area of research preferred by the researcher in the structural health monitoring. The wireless sensing unit is designed and developed for reliable communication of response measurements as well as for power efficiency. The extraction of signals for structural health monitoring is an important step and so a short illustration of wireless sensors is provided in this section**.** In [43], the authors had illustrated a brief review on the experiences of the investigators working on wireless sensors as well as sensor networks for maintaining records on structural performance and health. The structural monitoring systems involving wireless technology is a popular area of research because of its positive criteria such as low cost instillation since wiring is not essential between the sensors and data acquisition device. The stated methodology plays superior roles in the processing of structural response data since it is implemented to validate data involving signs of damages associated with the structure. The analysis on results related to international cooperative analysis on smart wireless sensors was carried out by [44]. In the initial phase of this research, the concept associated with the smart wireless sensor is investigated. In the next phase, the subsystems associated to the smart wireless sensor are illustrated. Apart from that, available wireless sensor platforms manufactured in the research companies are investigated. In the final phase, three smart wireless SHM technologies were demonstrated by the investigators which has vast scope in the identification of structural damages. In [45], the reseacher developed a novel wireless sensing unit device for intelligence monitoring of remote systems. The sensor is utilized for bridge as well as infrastructure SHM. The design is carried out based on cost reduction and superior performance. In [46], the authors illustrated an innovative wireless sensing device for structural health monitoring of historic landmarks. This technique minimizes the problems connected to the wired monitoring systems. In [47], the authors illustrated a design of an innovative low powered academic wireless sensor model for the monitoring of the structure. The effectiveness of the sensor design is assured via performance analysis. In [48], the researchers suggested a wireless sensing model incorporated with multitasking capabilities using recently widely available incorporated system components.

Table 2: Comparison of methodologies

Fuzzy Logic	Neural Network	Genetic Algorithm
Fuzzy logic is superiorly capable of handling uncertainties and complexities in structure with ease and effectively.	In case of arbitrary complex functions situations, neural network is considered to be as generalized function predictors.	Genetic algorithm is considered to be a powerful tool when implemented to solve combinatorial problems and so is advantageous in case of structural health monitoring.
Fuzzy logic is simple, easy to implement and robust in nature.	Neural Network has the capabilities of predicting the structural health during the operation without the necessity of interruption or the termination of the process.	Genetic algorithm is capable of determining the location of damage more precisely and computes the magnitude of damage more significantly.

5. Involvement of intelligent methodologies in structural health monitoring

5.1. Genetic algorithm

Genetic algorithm (GA) is advantageous because it offers easy implementation due to the fact that only an objective function is necessary and derivatives or other supplementary details are not required. So this attributes of GA makes it strong and most trust worthy in the field of structural optimization [49, 50]. The technique of GA have been used in model updating and detection of damage by few researchers in recent studies [51]. The technique of optimization utilizing the concept of genetics which is considered to be a zone of great interest among researchers investigating on the localized damage alteration as a function of **sets** involving modal [52]. In [53], the author had presented a technique of real number encoded genetic algorithm which is utilized in analysis of damages in structure utilizing method of objective function minimization. On the basis of available vibration data 3 separate attributes to be mentions as frequency variation, mode dimension variation and combination of this two

are taken into consideration. The direct collation is done by the objective function by the changes in the measured vibration extracts via damaged as well as undamaged structure before and after update**.** In their work, they used laboratory certified cantilever beam and frame to verify the efficiency of real-coded GA for the verification of structural damage. In [54], the researchers implemented the methodology of genetic algorithm in order to identify damages in vibrating structures. They utilized the concept of residual force matrix in combination of optimal methods for the identification of damages. The mentioned methodology is verified by the use of a doubled centered crossover dual embedded genetic algorithm complied with selected tournament for reproducing population. The optimal solution through GA is computed by utilizing C program for generating programs which utilizes the concept of eigenvalue forecasting code at par with sub program termed as force function residue. The methodology involves an appropriate selection of crossover as well as mutation operators. Considering the samples of 40 individuals with crossover probability 1.0, mutation probability 0.001., and each structural parameter β_i was illustrated as a 10 bit binary number having variable limits from 0 to 1. The main complexity involved is in obtaining the minimal force residue. Also the function related to the fitness is selected in the current matter termed to be as inverse function is illustrated as:

$$M = \frac{C_1}{C_2 + f(\beta_1, \beta_2, \ldots \beta_n)} \quad (1)$$

C_1 and C_2 are valued with number 1 and considered to be constants. The identification of damage parameters for various structures is done from associated residual vectors minimization problems that are handled by genetic algorithms. This technique explored the excellent bonding with those chosen for the mechanical simulation of these damaged structures. Tso and Shen [55] have used residual forced vectors in identification process based on neural network to display the properties of good location and severity assessment. In [13], the author had presented the concept of vectors associated with residual force in order to point out an objective function for a technique of optimization, which is imposed using genetic algorithm. The main intention involving the detection of structural damage is the formulation of parameters based on objective function. The formulation rules of objective function should include the process of obtaining the maximum value during the process of evaluation taking into consideration the true parameters. This is a matter of importance as the technique of optimization like the concept of genetic algorithm may be implemented to extract the parametric values using iterative methodology by choosing values for maximizing the objective function. The structural situation about the damaged states incurred can be established using the procedure of optimizing to get the solution values of the parameters. For addressing the identification of damage at par with constraint optimization related to the concept of GA, it is utter essential to broaden up the objective function. The suitable objective, which is required to be maximized by using the general rules of residual forces and taking into consideration several practical aspects, is given as follows:

$$J_{obj}(\beta_1 \ldots \beta_i \ldots) = \frac{C_1}{\sum_{f=1}^{r} R_f^*(\beta_1^* \ldots \beta_1^* \ldots)^T R_f^*(\beta_1^* \ldots \beta_1^* \ldots) + C_2} \quad (2)$$

where C_1 implies a constant used to control the value of the objective function C_2represents a constant used to construct a well established function for the ideal case and r represents the number of modes taken into consideration which is determined experimentally or used in simulated examples. In [56], the authors had used the approach of wave propagation having a synergistic combination with genetic algorithm in order to identify the damages in beam structures. In [57], the authors has proposed a technique of training a neural network taking into consideration the input parameters as strain and frequency whereas the output is taken to be the damage level. The neural network parameters are selected using GA. In [58], the authors had presented a GA-based methodology for the extraction of crack zone as well as aluminum made beam depth. The location of crack and depth in a cantilever beam is identified and constructed as an optimization problem. The optimal location and depth is found out by using binary and continuous genetic algorithms by implementing cost reduction scheme on the basis of extracted and computed changes in frequencies. Initially the chromosomes are arranged on the basis of their fitness. The selection probability P_n for each chromosomes considering rank n is:

$$\frac{N_{keep} - n + 1}{\sum_{n-1}^{N_{keep}} n} \quad (3)$$

Single point crossover is used for binary GA. In case of continuous genetic algorithm, the used variable value considered is unit offspring. P_{new} is considered to be the outcome of binary associated variable illustrated as:

$$P_{new} = \beta P_{mn} + (1 - \beta) P_{dn} \quad (4)$$

The interval [0, 1] is selected for β and is termed to be a random value. Also, P_{mn} is considered to be the nth mother chromosome variable. Finally, P_{dn} is nth variable associated with father chromosome. The methodology of genetic algorithm is implemented in this investigation for the identification of beam cracks. The identification methodology associated with the approach use unit and continuous GA for the identification of depth and crack. In [59], the researcher briefly defined a technique for damage identification of structure based on micro GA which utilizes test data involving noises and incompleteness. The computation of modal data is carried out by installing some sensors. In the primary phase, the expansion of mode shape data is executed in phase with all DOF related to the finite element model. The next step is to implement elemental energy quotient difference in order to investigate the damage location approximately. The final step includes the utilization of a micro-genetic algorithm for the quantification of the damage by minimizing the errors between the computed data and numerical outcomes. Ruotolo and Surface [60] used the methodology of genetic algorithms to identify structural damage in structure by the modal residuals optimization. In [61], the author had presented a combined methodology of genetic algorithm and finite element method for detecting and locating damage. In [62], the researchers had revealed a methodology based on nondestructive global damage detection and assessment computation associated with frequency changes and mode shapes of vibration of a structural system. This is an elementary level methodology based on finite-element model.

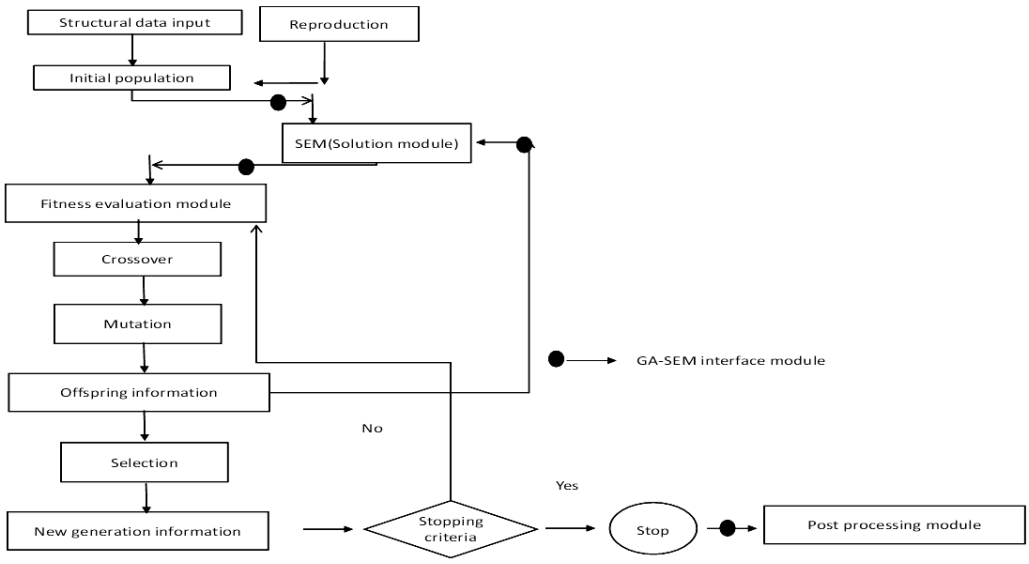


Figure 1: Flow chart of genetic search

Continuum damage mechanics suggest that the representation of damage should be done using a reduction factor of the element bending stiffness. In order to localize the area of damage in structures, an optimization approach with non-classical touch involving the implementation of genetic algorithm is laid down. Innumerable cases of damage instances has proved the efficiency of this method when applied to simulated beams and also by use of direct data from experiment which is extracted from the vibration tests of a beam. The proposed technique of genetic algorithm generates an appropriate location of damage and extreme detection while offering number of benefits in comparison to classical methods. In [63], the authors had presented an effective procedure involving optimization technique in order to locate multiple damage in structures. One indication that the structure is under the grip of damage is the natural frequency variation of the structure. The extraction of information on the basis of natural frequencies is accomplished by using a finite element analysis (FEA). A novel technique involving tweaked genetic algorithm(MGA) having novel operators termed as health as well simulation components are illustrated in significant way for detecting range and location of damages. The efficient combination of MGA and the ECBI that result in a effective robust tool for detecting structural damage is validated from the numerical results for different cases of damage with different sample. This methodology can be used to detect the real damage sites and extent thus decreasing the total number of FEA in comparison to the available tools like SGA-MDLAC. In [64], the authors had suggested a technique to compute the extent of damage as well as precise location using the concept of natural frequency shifts. The combination utilizes real genetic coded algorithm as an optimization tool to verify natural frequency with those extracted by using the technique of parametric finite element code via modified MDLAC index. A Binary-coded genetic algorithms (BCGA) concept was implemented for the detection of structural damage associated with GA technique [65]. In recent days, the trend of studies in the field of real-coded genetic algorithms (RCGA) is on the up rise [66, 67]. This approach works better with variables that are continuous and there is less requirement of storage than BCGA. An improvised GA technique for the detection of frame structure damages was presented by [68]. In their work they suggested discrete values to illustrate the damage. This methodology is composed of initial population of heuristic generation, two dynamically changing fitness functions on the basis of modal data and two appropriate mutation operators. In [69], the author suggested an algorithm in two stages in order to sense structural damage. The first step is to define an search algorithm in order to search for expected damages. In second case, the algorithm scanning is accomplished in order to uplift the level of knowledge. GA is utilized in all the cases. In [70], the authors had proposed a hybrid type real-coded genetic algorithm for the detection of structural damage. The objective function is incremented with a damage penalization in order to discard fake damage detection caused by experimental disturbance or error in computation. The algorithm verification is carried out using a three dimensional space frame structure taking onto consideration both single and multiple damage cases. This research reveals that the genetic algorithm approach used gives more accurate results than conventional optimization methods. The successful detection of real damage was made possible while neglecting the false damage detection. This methodology is widely used for large or complex damage criteria and preferred over traditional optimization techniques which fails to extract the global optimum. A hybrid methodology involving optic fiber sensor and GA was developed by [71] for structural health monitoring. The methodology involving random generation is utilized to develop the preliminary population having three individuals. The researchers have utilized the concept of objective function which is illustrated by addition of squared differences associated with the beam eigenvalues extracted obtained by a sensor termed as POF. Compilation of the frequency associated with the modal includes randomly generated individuals related to the initial population. The objective function is given as:

$$f(m,d) = \sum_{i=1}^{i=3} (f_{data}^{i} - f_{exp}^{i})^2 \quad (5)$$

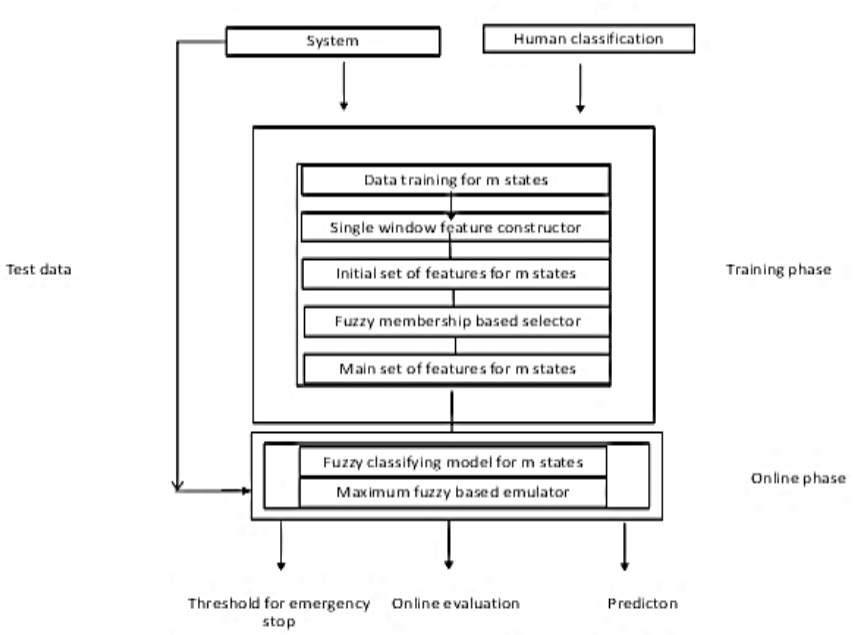


Figure 2: Fuzzy filter approach

objective function is carried out so that solution variables having minimal value will have the higher fitness value. GA is utilized in this investigation which is a novel approach. The one with the strongest feature is permitted to shift to the forthcoming generation. The approach of reimplementation associated with the chromosomes are embedded into the new population related to the forthcoming generation thus not permitting any variations in the GA. The approach of joint general GA along with sensor made of plastic optic fibre are depicted for the health monitoring of structures. The methodology utilized in this research can be implemented for identifying damages in structures made of concrete as well as composite materials. The identification of delamination in composite beams as well as joint structures utilizing an efficient identification concept was illustrated by [72]. The predicting technology based on modal analysis combined with damaged spectral element for extracting the structural response damage associated with the frequency domain is proposed. A novel type of curved genetic algorithm (GA) associated with a with finite element code for automatic is utilized for structural damage identification. The techniques involved in genetic search is illustrated in Figure 1. A technique for inverse problem is suggested for the detection and localization of damage in composite beam like structures by [73]. The developed methodology utilizes Genetic Algorithm (GA) as well as Particle Swarm Optimization (PSO) techniques. The basis of the objective function is the computation of natural frequencies and Modal Assurance Criterion. In [74], the author suggested a hybrid methodology based on of standard GA in order to improvise the stability associated with the EM algorithm involved in the findings of clusters optimizing number and also the parametric values thereby expanding the damage classified results. The importance of the illustrated approach is proved by utilizing real time sets of data extracted from the Z-24 Bridge situated in Switzerland. A varied feature GA was illustrated by [75] in order to detect the accurate locations as well as extents of the imparted minimal damages in the structure. The proposed methodology detects the damages to the suitable extent.

5.2. Fuzzy logic

In 1965, The concept of fuzzy theory was proposed by Zadeh [76]. Fuzzy algorithm is capable of processing simulated thoughts and judging capabilities of humans. One of the essential attributes of fuzzy control rule is its linguistic criteria which can be tweaked very easily and can be understood with ease**.** The nonlinear mapping can be handled effectively by the utilization of fuzzy set theory as well as fuzzy logic. The main attributes of fuzzy logic system is to deal uncertainty easily and with more flexibility thus making it popular among researchers. Fuzzy logic manages the oddly situations in a suitable way, also it is very sensitive to information such as it is possible for a peak may to get established as fault in the same time without creating problems [77]. Regarding the in-between laid dependencies, it is proved that the concept of fuzzy approaches results in superior outcomes [78, 79]. One aspect that makes fuzzy system a convenient tool for analysis is that it has convincingly less number of rules when compared to classical expert system [80]. Fuzzy logic technique, has astonishing capabilities of damage detection in structures along with ambiguous data processing capabilities [81, 82]. In [83], the authors had shown an intelligent approach to alter the human procedure with a computer by characterizing a system based on fuzzy concept by utilizing decision approach based on human attributes. The feature of developed technology is that it will grant affordable way of encoding methodology related to the rules set. The designing of fuzzy expert system is laid down on the basis of FE model associated with simple beam. Also, it is successful in validating reliable source of damage detection for every scenario of damage analysis. There is 100% accuracy in the detection of damages by this system and no fake positives or negatives on the shapes of the mode obtained by impact testing and or by using vibrometer procedure that are non-contact. The analysis reveals

that the expert system involving fuzzy portrayed in this paper displays a decisive and precise results in recognition of various cases of damage. The performance of thresholding based methods is overshadowed by the performance of this system mainly in the case of multiple damage. The elaborately defined example in this paper suggest that the thresholding based technique is capable of detecting only one of the two damaged areas but the expert fuzzy system proposed is well enough to detect both of the damaged areas. The decision process can be processed very easily by installing the damage detection system as an expert system. The intervention of fuzzy logic results in comfortable solution of the unreliability and denial within the rules set and while taking any decision. In [84], the authors had proposed a technique of conditioning monitoring and approach based on classification using fuzzy-filtering. There is a necessity of defining various models for differentiating different states of operation in a machine. This problem is discarded by model based on qualitative methods and signal based methods due to the involvement of innovative concept principle of modeling. In their work, the technique of qualitative model associated with fuzzy concept and statistic stated signal properties is combined synergistically. A conditioning monitoring system is designed on the basis of important signals, alternate measurements experimental data as well as the classifying knowledge of the system which is required to be verified. The concept behind the fuzzy filter is stated by a flow chart illustrated with Figure 2. The filtering methodology is illustrated relying on the concept of qualitative model based technique utilizing fuzzy concept in combination with statistics concept associated with signal based methods**.** The mentioned methodology gives the raw data conversion to a novel presentation for approaching the classification. In [85], the authors illustrated a technique for the validation of overlapping health fuzzy sets by utilizing presumed distributions of statistics associated with damage attributes. The superior impacts of implementing progressive stages of damages to be mention as crack, defect as well as flaw for laying down a systematic investigation of damage using the concept of supervised as well as unsupervised learning was mentioned by [86]. In [87], the authors implemented the methodology of fuzzy similarity for monitoring structural health utilizing damage pattern recognition. The health features are illustrated by fuzzy sets that are developed by utilizing the methodology of healthy observations. The representation of healthy as well as damaged states are accomplished by fuzzy membership functions which is utilized as a combination with piecewise linear functions. The model steel bridge is consider as a case study for identifying damage pattern in their work. The precise estimation of damage patterns is achieved by regularizing this methodology. The fuzzy set associate with health which is a left shouldered triangular membership function is illustrated as:

$$\mu_H(\lambda) = \begin{cases} 1, \lambda \le \bar{\lambda}_H \\ \frac{(\lambda H^* + \sigma_H) - \lambda}{\sigma_H}, \lambda H^* \le \lambda \le \lambda H^* + \sigma_H \\ 0, \lambda \ge \lambda H^* + \sigma_H \end{cases} \quad (6)$$

$\mu_H(\lambda))$ is stated as membership linked to the fuzzy set illustrating a good forma with a mean of λH^* having a spread of σ_{λ_H}. The membership function with good health illustrate that healthy fuzzy set considers a data interval which are completely leaned towards healthy outcome. The technique of pattern recognition associated with damaged states with unknown capability sets is implemented utilizing the concept of fuzzy similarity as well as maximum convex degree. The methodology depicted in this investigation can precisely detect the damages in the structure. The proposed technique is nonspecific for any SHM system due to the fact that it is independent of the damage feature. An advanced methodology was illustrated by [88] for the identification of location as well as damages extremity in the area of stochastic structures relied on ARMA parametric model combined with fuzzy classification. The method suggested in this paper has a high significance as it uses fuzzy viewpoint on stochastic structural damage analysis. In this case, the feature vector is considered to be ARMA parameters. The ability of this technique is also minimize the stage at par with the optimization which is utilize to extract the values related to the membership functions of fuzzy technique by implanting changes of predicted ARMA values associated with membership functions. In this current investigation, the fuzzy classified inputs are ARMA model's parameters and the outputs are specified as elemental damaged numeric as well as the level of extremity. Considering all features, the antecedent fuzzy sets (membership function) is illustrated as:

$$\alpha\left(\mathrm{f_i^j}\right) = \exp[0.5]\left(\frac{\mathrm{f_i^j - m_i^j}}{\sigma_\mathrm{i}^\mathrm{j}}\right)^2 \quad (7)$$

where, $\alpha(f_i^{\,j})$ = ith membership function of jth feature, $f_i^{\,j}$ = absolute value, m_i^j = related mean value (midpoint) and $\sigma_i^{\,j}$ = variance of the jth feature in ith class. In [89], the authors suggested a noteworthy methodology on the basis of fuzzy clustering combined with computed frequency response function samples minimized via principal component projection in order to investigate damage detection in truss bridge model using vibration based technique. The preliminary data utilized in this investigation are the FRFs associated with the healthy as well as the damaged structure. The normalization of FRF data is carried out to eliminate the effects caused by the environmental and operational variability. The analysis by principal component at par with kernel principal component investigation technique are laid down for handling the compression of data. In extraction methodology of damage features, the principal components of median values are used. Finally, fuzzy c-means(FCM) clustering algorithm is applied for the categorization of structural damage detection features. In [90], the authors had suggested an approach of novel binary phase detection of damages in structures utilizing the theory of fuzzy-neuro methodology and techniques of data fusion. This method is widely used for the identification of damage considering the condition that computed sample is vast and involves uncertainties. Primal phase of damage detection involves the collection of modal attributes of structure from the structural vibration responses that are embedded as an input to the FNN. The output values associated with the FNN are defuzzified for delivering a rough evaluation damages in structure**.** In the second phase, FNN model reveals three output values which are inputted to the center of data fusion in order to execute fusion computation. The result is then filtered using threshold function in order to carry out final fusion decision. Thus a superior quality damage assessment methodology is achieved. The innovative approach is implemented in 7 DOF building for the validation of damage identification of structure.

The results are effective, reliable and satisfactory. In [91], the researchers had proposed an modernized approach to detect damage in structures using a new algorithm based on the fuzzy cognitive map (FCM). In this research, damage in structures is modeled with the help of continuum mechanics approach. The technique of modeling using finite element method for cantilevers is utilized in order to compute the variation of first six beam frequencies due to structural damage. In [92], the researchers illustrated the concept of structural damage detection based on data fusion as well as semi-supervised fuzzy C-means clustering. This method performs superiorly when compared with other methods for detecting damages in structures. The experimental results associated with a benchmark model validates the effectiveness of the proposed methodology. In [93], the authors illustrated a fuzzy inference model for the extraction of building damage hazard, normally associated with the non-technical building houses considering special sort of earthquake event. The laid down model consists of 3 level model consisting of fuzzy rules which utilizes numerable data associated to the damaged buildings obtained from the impingement of earthquake occurred in 2013 at Indonesia. The important contribution associated with this research is vested towards the development of the function as well as membership related to fuzzy taking into consideration the variable of each determinant associated to building house damage warning as well as three levels of fuzzy inference procedure for extracting building house damage hazard as an earthquake occurrence. A system based on decision making and associated with an adaptive network-based fuzzy inference system (ANFIS) model was proposed for the extraction of damage level of building by [94]. The suggested methodology was tested over earthquake data set of Bam city in Iran. The proposed methodology was accurate up to 76.36% for the detection of damage degree of building. A technique for damage identification in buildings using the approach of superior dimension remote-sensing imaging combined with fuzzy reasoning concept was presented by [95].

5.3. Concept of Neural Network

A neural network consists of inter connected neurons and layers and it is one of the most important means of computing weights which are relied on the connections illustrated as learning algorithms. In [96], the authors has vividly described the concept of common anatomy of neural network by sub dividing it into seven important zone as follows: i) processing unit set, ii) activation set of processing unit, iii) calculation of processing unit output using a function, iv) connectivity pattern alongside the processing unit, v) implementation of propagation rule, vi) instillation of activation function and vii) confirmation of learning rule. In [97], the author had investigated the primary research stage inclined at the development of automatic monitoring methods for the detection of damage in structure. In their study, the technique of self-organization and learning capabilities of neural network are explored for the assessment of structural damage. The concept involved is the training approach of neural network to identify the undamaged structure behavior alongside the behavior of the structure considering numerous possibility of damage states. The detection of any existing damage can be made possible by impinging the training network for the computation of responses from structure. This concept was implemented to a simple structure to get a motivated results. The initial outcome of results verifies that the neural network are capable enough to learn the nature of damaged and undamaged structures and thus can be used to identify the damaged member and the extent of the damage by the approach of patterns in the frequency response of the structure. In [98], the author utilized the concept of counter propagation neural network for revealing the beam damages, also damages in a frame as well as support movements of a beam along axial direction**.** The research involves a wide range of diagnostic parameters, along with static displacements, natural frequencies, mode shapes and other mode shapes parameters. The fetching results are quite interesting showing that these parameters can be utilized as a diagnostic parameters for artificial neural network applied to structural engineering. In [99], the researchers developed an artificial neural network technique associated with system identification concepts extracting the characteristics of composite structures damage. The methodology implemented in this research consider a training phase as well as recognition phase. The different types of damage considered to be as patterns are associated with structures candidate model in the training phase. The area and the extremeness of the damage decides the arrangements of patterns into pattern classes. In order to extract the transfer functions, the identification of system are carried out which is an important attributes associated with the structural systems. The extracted transfer functions is used as input patterns for training which are transferred to multi-layer perceptron. Numerical validation proves the efficiency of the developed technique. In [100], the authors implemented an architecture on the basis of multiple layer perceptron in combination with the learning algorithm based on backpropagation for modeling a bridge by utilizing data from damaged states. By using the methodology of general structure analyzer, the patterns of the training were laid down by predicting the zones of structural damages. By creating a comparison between output and the performance data that are created algorithmically without considering it in the training process, the working of the network was displayed. The problems associated to the working of the network is thoroughly investigated. For the estimation of structural damage it can be accomplished from various investigation done that the perceptron model is completely applicable. For the described case, the functioning of the architecture of an individual layer is not more sophisticated than the performance with double hidden layers. The investigation done on the calculated data based on the implications of engineering is considered by the evidence that for the identification recognition task barely a certain area in the structure required to prepare the network. In [101], the authors conducted an experiment five storey frame shake table and validated that there are problems leading to damage states occurred during the event of implementation of neural concept that is trained utilizing trained samples obtained by using FEM. The obtained results proves the superiority of the technique. In [102], the authors illustrated a novel network associated with proposition for the identification of damages in structure. In the preliminary part, the technique initiates with identification of system utilizing networks on the basis of neural concept recognition in order to extract the damaged as well as undamaged states. The next phase deals with the structural damage identification using neural concept in combination with damage identification network for the area detection as well as severity of the structural damage. Considering n hidden layer in the network having output with k^{th} node:

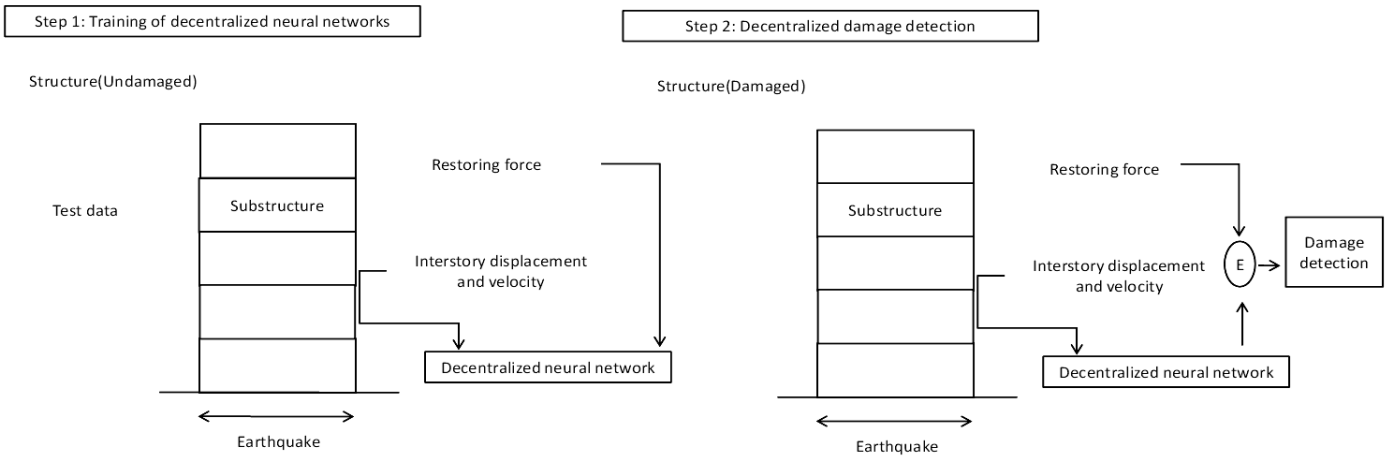


Figure 3: Decentralized damaged detection

(8)

H_{nj} is the illustrated output at par with j^{th} node having n^{th} hidden layer. Again, θ_{ok} is the threshold parameter related to k^{th} node of the output. Also, $W_{hnj,ok}$ is termed to be the weight of the j^{th} node embedded in n^{th} hidden layer of the k^{th} node of the output. The activation function is illustrated by f. The approaches implemented in this research is highly effective in structures involving nonlinearities, as neural concept can handle linearity and non-linearity with same technique. An innovative methodology of decentralized information structures related to the neural network concept was proposed by [103]. An identification technique involving parametric damage with decentralized pathway is implemented to the structures with multi degrees of freedom. In preliminary stage, detection of damages associated with substructures of a multi degree of freedom system is implemented and is applied by the method of neural networks. The training of the decentralized neural network required for the detection of analogous substructure is carried out based on displacements as well as velocity computations obtained from a good health substructure system as well as from the substructure restoring force. The differences in forces between the damaged substructures as well as the undamaged substructures is computed by means of trained decentralized identification neural concept. The figure 3 demonstrates the block diagram related to decentralized damage detection with the application of neural networks. The significance quality level of this methodology was inspected using numerical simulations. In this current research, the investigator has shown the effectivity of this method by demonstrating it as a practical tool for damage detection which can be utilized to structure-unknown smart civil structures. In [104], the authors laid down the technique of modular neural network approach in for the extraction of crack zone as well as depth associated with a cantilever beam. In the mentioned investigation, principal component analysis network is implemented for size mitigation as well as for unearth the concealed structure associated with the data set. In the adopted modular network approach, the crack location is estimated using one neural network whereas the other is used for computing depth of crack. In [105], the researchers presented an algorithm for the detection of damage in beam like structures with prediction of damaged ranges and zone by implementing the novel methodology of global and local vibration dependent analysis samples as input in neural concepts. Performance sensitivity analysis give rise to the essential features for damage detection. Various input-output sets are incorporated to various ANNs. The verification of input robust features used in the methodology as well as the actuation of the experimental uncertainties are done by generating artificial random noise and adding noise free data during the training procedure of the ANNs. The procedure of sensitivity analysis is carried out on the extracted features by using different modes of vibration taking into account the effect of the area of damage as well as severity before introducing them to ANNs. The various steps involved in the investigation starting from vibration based analysis for feature extraction in order to use as an input for the ANNs, noise level on these features, the methodology of the experiments, measuring instruments to the effectiveness verification at different vibration modes plays a significant role in extremity and location forecasting of the damage in beam like structures. In [106], the authors had proposed a methodology for damage detection using neural networks in association with modal data which impart the parameters of input to the modeling errors. In their study, to implement the methodology practically, the estimation of damage in bridge structures is demonstrated utilizing the technique of ambient vibration samples generated by traffic loadings. By initiating two numerical processes, the effectivity of the laid down concept is validated. The next step involves tests in laboratory as well as field tests on the structure of the bridge. The results are promising and effective which validates the efficiency of the neural network technique for damage identification, locations detection and severity analysis. In [107], the researchers illustrated a MLP neural network relied concept for damage estimation of structure related to the truss bridge joints, which is obtained from modal specification of the structure. Two mathematical investigation samples are illustrated to confirm the superiority of neural network concept. In [108], the authors presented an innovative approach of neural network based structural damage detection. The proposed methodology can be subdivided into categories. The first step includes system identification techniques that uses neural system identification networks (NSINs) to track the undamaged and damaged states of a structural system. The second step is associated to structural damage detection that uses trained NSINs to create free vibration responses taking into

consideration the same initial condition or impulsive force. The level of changes can be made available by comparing the periods and amplitudes of the free vibration responses of the damaged and undamaged states. In [109], the authors demonstrated a statistics based neural network approach by considering finite element modelling error as well as measurement noise for the detection of structural damages. The investigation on numerical as well as experimental outcomes reveals that the superiority of statistical neural network concepts approach in comparison to the normal neural network approach generates more precise identification of damages in structure. Once the model training of neural network is over, the testing samples are then used to identify the locations and severities of any damages. In [110], the authors investigated a damage detection procedure with the help of identification of pattern methodology associated with vibration. The strategy is evaluated utilizing a FEM of a suspension bridge which is a century old. The response spectra give rise to feature vectors which was installed to binary unsupervised neural networks for evaluation. Investigation reveals that the tweaking of the sensitivity of the neural networks is accomplished for achieving a superior rate of damage detection along with the existence of noisy signals. In [111], the authors suggested the methodology of novelty detection and damage computation. In this process, signal parameters obtained from reference structure are used as a data set. Also soft computing methods are used to notify about the damage visualization, location and its type. The results obtained from the investigation shows that the suggested methodology makes automation of structure testing possible and can be applied to SHM systems. In this investigation, a two-stage SHM algorithm was developed which can predict structure damaged height and width. Neural Network (NN) trained novelty detection is used for the first level identification. The second level of identification is conducted using the NN trained for damage prediction. In [112], the authors proposed a damage detection methodology on the basis of computational method to identify damage in structures using neural networks. In this process, the training of the neural network is done for the recognition of the pattern of static value associated with the undamaged structure as well as of the structure having several possible damage extent and location which were modeled as random states. The damages in the cantilever beam was identified using this methodology. This study reveals the fact that the proposed methods using strain is more effective than the displacement method. In [113], the researchers presented an approach for the detection of small damage in structures using ANN technique with progressive substructure zooming. In this work, a combination of substructure technique along with a multi-stage ANN models in order to detect the area and extremity of damage. The input for ANN are considered to be modal parameters such as frequencies and mode shapes. The higher performance of the technique is validated using a two-span continuous concrete slab structure and a three-story portal frame. The experimental results shows that this approach is able to identify all the simulated damages in the structure. An identification methodology for the extraction of damages in structure by implementing the technique of parametric identification combined with iterative neural network (NN) concept was proposed by [114]. A variable back-propagation learning algorithm is proposed for tackling possible saturation of the sigmoid function as well as to run the training process. The structural parameters are predicted by utilizing the trained NN model which is implemented in measured dynamic attributes. The extracted structural parameters is then impinged in the finite element model for computing the dynamic characteristics. The obtained structural parameters is utilized for the identification of the location as well for examine how severe is the structural damage. In [115], the authors illustrated a particle swarm optimization-based methodology in order to train the NN (NN-PSO). The suggested approach is well capable of handling the problem of predicting failures in structure associated with multistoried reinforced concrete buildings. The superiority of the methodology is validated with the help of experimental results. In [116], the researchers developed a damage detection technique that uses 3 alternate neural networks. In order to analyze the model, a tiny dataset having ground-level image data of damaged buildings was extracted. Experimental results demonstrated the efficiency of the proposed method. In [117], the authors proposed a novel approach for the detection of structural damage as well as estimation utilizing incomplete noisy modal data and artificial neural network (ANN). A feed-forward back propagation network is suggested for the estimation of the structural damage location and severity. In [118], the author illustrated an innovative, fast and precise identification of damages in structure using 1D Convolutional Neural Network. The illustrated technique performs vibration based damage identification as well as localization of the damage in real-time and generates superior outcomes.

5.4. Hybrid Techniques

Fuzzy logic, artificial neural network and genetic algorithm are the three main important categories of intelligent techniques that tries to implement the perception of human intelligence using different methodologies. There is positive and negative attributes of each methodology that differentiates among each other. The positivity of neural network is that it can learn accurately from data and establish highly nonlinear mapping among various platforms but it is concealed to the users and it has very low reasoning capabilities. This drawbacks can be compensated by the techniques of IF-THEN rules and logical inference from the concept of fuzzy logic. The linguistic models is the main approach of fuzzy logic that uses the same decisions making capabilities of human beings as in real life. Linguistic models inherits high reasoning and nonlinear mapping capabilities. But the constraint embedded to fuzzy systems is that it cannot cope up with the ability of learning and cannot shape up according to a new ambient. The information processing systems like classifiers and controllers are the main objectives of artificial neural network and fuzzy logic unlike genetic algorithm. The genetic algorithm has a higher degree of parallel searching capabilities that enables it to search multi-peak spaces. Another attributes of genetic algorithm is that it is not dependent on continuity of the parameter space. Apart from the positivity of genetic algorithm, the drawback of genetic algorithm is that it cannot be used alone as modeling techniques. It is very much efficient when it is made to work in parallel with other modeling techniques. The combination of all three intelligent concepts will yield maximum favorable outcome and vastly accepted as a hybrid methodologies [119]. A learning technique at par with the neural network is developed by [120] which is relied on the genetic fuzzy combination. The approach is to facilitate training to the radial basis function associated with the neural network. The unique approach is a blend of genetic algorithm with

fuzzy logic in order to tweak the centers as well as widths linked to the RBFNN. By utilizing the technique of linear least square method, the optimization of neural network connection weights is achieved. The output linked to the RBF fuzzy neural network is stated as:

$$Y_{pi} = \sum_{k=1}^{3^6} W_{ik}\bar{\alpha}_{pk},\ i=1,2 \quad (9)$$

The matrix α depicts the firing strength of all rules. Also, W is illustrated as the output linked weight matrix. Again, Y is considered to be output matrix. The equation depicting the fitness function with genetic algorithm is:

$$F_{i=}\begin{cases}\bar{J}+2\sigma-J_i, if\ J_i<\bar{J}+2\sigma, i=1,2,\dots 200\\ 0, otherwise\end{cases} \quad (10)$$

The mean value associate linked to the objective function is defined by $\bar{J}$. Also, σ is considered to be standard deviation and J_i is the parameter associated with objective function. A hybrid neuro-genetic algorithm for evaluating the damages in the structure is presented by [57]. Current investigation reveals that two hidden layer network was chose taking into consideration back propagation algorithm with momentum coefficients as training algorithm. The network parameters in this case which are number of neurons associated to hidden layer, learning as well as momentum coefficients are taken as variables. Also, genetic algorithm is implemented for choosing the appropriate sets of the parameter. Superiority of the developed algorithm is proved by implementing dual structures consisting of a frame and a beam. The equation representing the objective function linked to the optimization problem is:

$$E(X_{k,}W) = \frac{1}{p}\sum_{k=1}^{p}\sum_{m=1}^{N_n}(y_{k,m}-o_{k,m})^2 \quad (11)$$

the desired and obtained outputs for the m^{th}output and k^{th}training instance are $y_{k,m}$ and $o_{k,m}$ respectively. In [121], the authors presented a damage detection algorithm on the basis of neuro fuzzy hybrid system for the study of location as well as severity predictions of cracks associated to beam-like structures. In [122], the authors proposed an appropriate fuzzy inference system (FIS) which is also termed as an Adaptive Neuro-Fuzzy Inference System for the detection of damage level in roads. The extracted results display the effectiveness and preciseness of the Neuro-Fuzzy systems for road damage detection. In [123], the authors developed a semi-automated Mamdani based fuzzy decision-making system for the identification of damages in buildings by the utilization of textural features. For the improvisation of the level of automation in Mamdani fuzzy system, a genetic algorithm was utilized to find its optimized parameters associated with membership functions.

6. Conclusions

Structural health monitoring is a wide are of research and has different intentions depending on the various scenario. But approaches that need to be implemented have common and specific component classes at the levels mentioned as follows: a) Need for sensors, b) storage of data, c) data transmission technique, d) database management system in order to carry out feature extraction, e) development of model from the extracted data, f) past data analysis for knowledge and g) steps related to decision making on the basis of features that is identified in combination of known models. The design state of structural health monitoring is an important phase where consideration needs to be implemented in order to monitor the substructures related to uncertain or critical performance [124]. This critical survey revealed various positive attributes and capabilities of intelligent technique in the detection damaged structures associated to civil engineering. The genetic algorithm can generate good approximate solutions related to complex optimization problems. Fuzzy logic displayed its capabilities on handling the structural systems associated with complexities. The fuzzy methodology can handle uncertainty with ease. The neural network established its identity as it has above well capabilities in detecting cracks in structures. It also shows superior performance in detecting damage index of cases associated to the free decay of structures. It is to be noted down that hybrid techniques are more effective than other conventional artificial intelligent methods in structures as it induces the quality of individual intelligent techniques in a single methodology. The intelligent technique is considered to be the most popular tools for damage identification due to its precise, reliable and low-cost solution nature. The main intention of this survey paper is to provide an in-depth knowledge to the researcher about the application of various intelligent approaches in the area structural health monitoring. This research will impart a vast knowledge of intelligent tools among the researchers who wants to conduct analysis and monitoring of structures under various conditions. This survey paper will generate a brief idea among the researchers, the several ways of initiating research in the area of structural health monitoring.

References

[1] Richardson, M.H. Detection of damage in structure from changes in their dynamic (modal) properties---a survey. 1980; NUREG/CR-1431,Washington, DC: US Nuclear Regulatory Commission.

[2] Rytter, A. Vibration based inspection of civil engineering structures. PhD Dissertation. 1993; Department of Building Technology and Structural Engineering, Aalborg University, Denmark.

[3] Chang, P.C.;Flatau, A.; Liu, S.C. Review paper: health monitoring of civil infrastructure. Struct Health Monit, 2003; Volume 2, Issue No. 3, pp.257--67.

[4] Doebling, S.W.; Farrar, C.R.; Prime, M.B.; Shevitz, D.W. Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. 1996; Report LA-13070-MS, Los AlamosNationalLaboratory, Los Alamos, NM.

[5] Ubertini, F., Comanducci, G., Cavalagli, N., Vibration-based structural health monitoring of a historic bell-tower using output-only measurements and multivariate statistical analysis. 2016; Volume 15, Issue No.4, pp.438-457.

[6] Bhuiyan, M. Z. A., Wang, G., Wu, J., Cao, J., Liu, X., Wang, T., Dependable Structural Health Monitoring Using Wireless Sensor Networks, 2017; IEEE Transactions on Dependable and Secure Computing, Volume 14, Issue No. 4, pp. 363-376.

[7] Carden,E.P.;Fanning,P.Vibration based condition monitoring: A review, Structural Health Monitoring,2004; Volume 3, Issue No. 4, pp.355--377.

[8] Auweraer, H.V. International research projects on structural damage detection, Damage Assessment of Structures Key Engineering Materials, 2001; Volume 204, Issue No. 2, pp.97—112.

[9] Shi, Z.Y.;Law, S.S.; Zhang, L.M. Structural damage detection from modal strain energy change, Journal of Engineering Mechanics-ASCE , 2000; Volume 126, Issue No.12,pp.1216--1223.

[10] Gawronski, W.; Sawicki, J.T. Structural damage detection using modal norms, Journal of Sound and Vibration, 2000; Volume 220, Issue No. 1, pp.194-198.

[11] Rytter, A. Vibration based inspection of civil engineering structures, PhD Dissertation.Department of Building Technology and Structural Engineering, Aalborg University, Denmark; 1993.

[12] Tsou, P.; Shen, M.H.H. Structural damage detection and identification using neural networks, AIAA J, 1994; Volume 32, pp. 176-183.

[13] Mares, C.; Surace, C. An application of genetic algorithms to identify damage in elastic structures, J Sound Vib,1996; Volume 195, pp. 195-215.

[14] Fuat, E.; Oguzhan, H.; Ilker, T.; Hakan, K.Optimal design of planar and space structures with genetic algorithms, Compos.Struct., 2000; Volume 75, pp. 209-224.
[15] Koh, C.; Chen, Y.; Liaw, C. A hybrid computational strategy for identification of structural parameters, Computers and Structures, 2003; Volume 81, Issue No. 2, pp. 107-117.
[16] Friswell, M. I.; Penny, J. E. T.; Garvey, S. D. A combined genetic and eigensensitivity algorithm for the location of damage in structures, Compos.Struct., 1998; Volume 69, pp.547-556.
[17] Chou, J.; Ghaboussi, J. Genetic algorithm in structural damage detection, Compos.Struct., 2001; Volume 79, pp. 1335-1353.
[18] Chen, D.; Wang, W. J.Classification of wavelet map patterns using multi-layer neural networks for gear fault detection, Mechanical Systems and Signal Processing, 2002; Volume 16, Issue No. 4, pp. 695-704.
[19] Garg, A. K.; Roy Mahapatra, D.; Suresh, S.; Gopalakrishnan, S.; Omkar, S. N.Estimation of composite damage model parameters using spectral finite element and neural network, Composites Science and Technology, Volume 64, Issue No. 16, pp. 2477-2493.
[20] Ko, J. M.; Sun, Z. G.; Ni, Y. Q. Multi-stage identification scheme for detecting damage in cable-stayed KapShuiMun Bridge, Engineering Structures, 2002; Volume 24, Issue No. 7, pp.857-868.
[21] Gao, J.Introduction to intelligent information processing. Beijing: China Machine Press, 2004.
[22] Pedrycz, W.Computational intelligence. Boca Raton, FL: CRC Press, 1997.
[23] Ganguli, R.A fuzzy logic system for ground based structural health monitoring of a helicopter rotor using modal data, Journal of Intelligent Material Systems and Structures, Volume 12, Issue No. 6, pp. 397-407.
[24] Silva, S.D.; Dias, M.; Lopes, V.; Brennan, M.J.Structural damage detection by fuzzy clustering, Mechanical Systems and Signal Processing, 2008; Volume 22, Issue No. 7, pp. 1636-1649.
[25] Kong, F.; Chen, R. A combined method for triplex pump fault diagnosis based on wavelet transform, fuzzy logic and neuro-networks, Mechanical Systems and Signal Processing, 2004; Volume 18, Issue No. 1, pp. 161-168.
[26] Meyyappan, L.; Jose, M.;Dagli, C.; Silva, P.; Pottinger,H. Fuzzy-neuro system for bridge health monitoring, 22ndInternational Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003, Chicago, IL, 2003; pp. 8-13.
[27] Roy, N.; Ganguli, R. Filter design using radial basis function neural network and genetic algorithm for improved operational health monitoring, Applied Soft Computing, 2006; Volume 6, pp. 154-169.
[28] Lee, Y.S.; Chung, M.J.A study on crack detection using Eigen frequency test data, Computers &Structures, 2000; Volume 77, Issue No. 3, pp.327-342.
[29] Kim, J.T.; Ryu, Y.S.; Stubbs, N. Damage identification in beam-type structures: frequency-based method vs. mode-shape based method, Engineering Structures,2003; Volume 25, Issue No. 1, pp. 57-67.
[30] Deraemaeker, A.; Reynders, E,; De Roeck, G.; Kullaa, J.Vibration-based structural health monitoring using output-only measurements under changing environment, Mechanical Systems and Signal Processing, 2008; Volume 22, Issue 1, pp. 34-56.
[31] Khoo, L.M.; Mantena, P.R.; Jadhav, P. Structural damage assessment using vibration modal analysis, Structural Health Monitoring, 2004; Volume 3, Issue No. 2, pp. 177-194.
[32] Yan, A.M.; Kerschen, G.; De Boe, P.;Golinval, J.C. Structural damage diagnosis under varying environmental conditions---Part I: A linear analysis, Mechanical Systems and Signal Processing, 2005; Volume 19, pp. 847-864.
[33] Gelman, L.The new frequency response functions for structural health monitoring, Engineering Structures, 2010; Volume 32, Issue No. 12, pp. 3994-3999.
[34] Bandara, R. P.; Chan, T. H. T.; Thambiratnam, D.P. Structural damage detection method using frequency response functions, Structural Health Monitoring, 2014; Volume 13, Issue No. 4, pp. 418-429.
[35] Curadelli, R.O.; Riera, J.D.; Ambrosini, D.; Amani, M.G. Damage detection by means of structural damping identification, Engineering Structures, 2008; Volume 30, Issue No. 12, pp. 3497-3504.
[36] Jaishi, B.; Ren, W-X. Damage detection by finite element model updating using modal flexibility residual, Journal of Sound and Vibration, 2006; Volume 290, Issues No. 1–2, pp. 369-387.
[37] Huth, O.; Feltrin, G.; Maeck, J.; Kilic, N.; Motavalli, M. Damage Identification Using Modal Data: Experiences on a Prestressed Concrete Bridge, Journal of Structural Engineering, 2005; Volume 131, Issue No. 12, pp.1898-1910.
[38] Xia, Y.; Hao, H.; Brownjohn, J. M. W.; Xia, P.-Q. Damage identification of structures with uncertain frequency and mode shape data. Earthquake Engng. Struct. Dyn., 2002; Volume 31, pp.1053–1066.
[39] Furukawa, A.; Otsuka, H.; Kiyono, J. Structural Damage Detection Method Using Uncertain Frequency Response Functions, Computer-Aided Civil and Infrastructure Engineering, 2006; Volume 21, pp.292–305.
[40] Yan, J.; Cheng, L.; Wu, Z.Y.; Yam, L.H. Development in vibration-based structural damage detection technique, Mechanical Systems and Signal Processing, 2007; Volume 21, Issue No. 5, pp. 2198-2211.
[41] Hou, Z.; Noori, M.; Amand R. St.Wavelet-Based Approach for Structural Damage Detection, Journal of Engineering Mechanics, 2000; Volume 126, Issue No. 7, pp. 677-683.
[42] Escamilla-Ambrosio, P.J.; Liu, X.; Lieven, N.A.J.; Ramírez-Cortés, J.M. Wavelet-fuzzy logic approach to structural health monitoring, Fuzzy Information Processing Society (NAFIPS), 2011; Annual Meeting of the North American, El Paso, Texas, USA.
[43] Lynch, J. P.; Loh, K. J. A Summary Review of Wireless Sensors and Sensor Networks for Structural Health Monitoring, The Shock and Vibration Digest, 2006; Volume 38, Issue No. 2, pp. 91-128.
[44] Cho, S.; Yun, C-B.; Lynch, J. P.; Zimmerman, A. T.; Spencer Jr., B. F.; Nagayama, T. Smart Wireless Sensor Technology for Structural Health Monitoring of Civil Structures, Steel Structures, 2008; Volume 8, pp. 267-275.
[45] Aoki, S.; Fujino, Y.; Abe, M. Intelligent Bridge Maintenance System Using MEMS and Network Technology, In Smart Systems and NDE for Civil Infrastructures, San Diego, CA, Proceedings of the SPIE, 2003; Volume 5057, pp. 37–42.
[46] Casciati, F.; Faravelli, L.; Borghetti, F.; Fornasari, A. Tuning the Frequency Band of a Wireless Sensor Network, In Proceedings of the 4th International Workshop on Structural Health Monitoring, Stanford, 2003; pp.1185–1192.
[47] Ou, J.; Li, H.; Yu. Development and Performance of Wireless Sensor Network for Structural Health Monitoring," Smart Structures and Materials, , Proceedings of the SPIE, 2004; Volume 5391, pp.765–773.
[48] Wang, Y.; Lynch, J. P.; Law, K. H. Wireless Structural Sensors Using Reliable Communication Protocols for Data Acquisition and Interrogation, Proceedings of the 23rd International Modal Analysis Conference, 2005.
[49] Camp, C.; Pezeshk, S.; Cao, G. Optimized Design of Two-Dimensional Structures Using a Genetic Algorithm, Journal of Structural Engineering, 1998; Volume 124, Issue No. 5, pp. 551-559.
[50] Jenkins, W. M. Towards structural optimization via the genetic algorithm, Computers & Structures, 1991; Volume 40, Issue No. 5, pp. 1321-1327.
[51] Rajeev, S.; Krishnamoorthy, C. S. Discrete optimization on structures using genetic algorithms, J. Struct. Eng., 1992; Volume 118, Issue No. 5, pp. 1233-1250.
[52] Larson, C. B.; Zimmerman, D. C. Structural model refinement using a genetic algorithm approach, Proc., 11th Int. Modal Analysis Conf., Society for Experimental Mechanics, Kissimmee, Fla., 1993; pp. 1095-1101.
[53] Yong, X.; Hong, H. A Genetic Algorithm for Structural Damage Detection Based on Vibration Data. School of Civil and Structural Engineering, Nanyang Technological University.
[54] Rao, M. A.; Srinivas, J.; Murthy, B. S. N. Damage detection in vibrating bodies using genetic algorithms, Computers and Structures, 2004; Volume 82, pp. 963-968.
[55] Tso W.K.; Zhu, T. J. Design of torsionally unbalanced structural systems based on code provisions I: ductility demands, Earthquake Engineering & Structural Dynamics, 1992; Volume 21, Issue No. 7, pp. 609–627.
[56] Krawczuk, M. Application of spectral beam finite element with a crack iterative search technique for damage detection, Finite Element Anal. Des., 2002; Volume 38, pp. 537-548.
[57] Sahoo, B.; Maity, D. Damage assessment of structures using hybrid neuro-genetic algorithm, Appl. Soft Comput., 2007; Volume 7, pp. 89-104.
[58] Vakil-Bagmisheh, M-T.; Peimani, M.; Sadeghi, H.; Ettefagh, M. M. Crack detection in beam-like structures using genetic algorithms, Applied Soft Computing, 2008; Volume 8, pp. 1150-1160.
[59] Au, F. T. K.; Cheng, Y. S.; Tham, L. G.; Bai, Z. Z. Structural damage detection based on a micro-genetic algorithm using incomplete and noisy modal test data, , Journal of Sound and Vibration, Volume 259, Issue No. 5, pp. 1081-1094.
[60] Ruotolo R.; Surace C.Damage assessment of multiple cracked beams: numerical results and experimental validation, Journal of Sound and Vibration, 1997; Volume 206, Issue No. 4, pp. 567-588.
[61] Khatir, S.; Belaidi, I.; Serra, R.; Benaissa, B.; Saada, A. A.Genetic Algorithm Based Objective Functions Comparative Study for Damage Detection and Localization in Beam Structures, Journal of Physics: Conference Series, 2015; Volume 628.
[62] Perera, R.; Torres, R. Structural Damage Detection via Modal Data with Genetic Algorithms, Journal of Structural Engineering, 2006; Volume 132, Issue No. 9, pp. 1491-1501.
[63] Nobahari, M.; Seyedpoor, S.M. Structural damage detection using an efficient correlation-based index and a modified genetic algorithm, Mathematical and Computer Modelling, 2011; Volume 53, pp. 1798-1809.
[64] Gomes, H.M.; Silva, N.R.S.Some comparisons for damage detection on structures using genetic algorithms and modal sensitivity method, 2008; Volume 32, pp. 2216-2232.

[65] 65. Perera, R.; Manzano, C. Performance assessment of multicriteria damage identification genetic algorithms, Computers & Structures, 2009; Volume 87, Issue No. 1-2, pp. 120-127.
[66] 66. Hao, H.; Asce, M.; Xia, Y.Vibration-based damage detection of structures by genetic algorithm, Journal of Computing in Civil Engineering, 2002; Volume 16, Issue No. 3, pp. 222-229.
[67] He,R-S.; Hwang, S-F. Damage detection by an adaptive real-parameter simulated annealing genetic algorithm, Computers and Structures, 2006; Volume 84, Issue No. 31-32, pp. 2231-2243.
[68] Borges, C.; Barbosa,H.;Lemonge,A.A structural damage identification method based on genetic algorithm and vibrational data, International Journal for Numerical Methods in Engineering, 2007; Volume 69, pp. 2663-2686.
[69] Kouchmeshky, B.; Aquino, W.;Bongard,J.; Lipson, H. Co-evolutionary algorithm for structural damage identification using minimal physical testing, International Journal for Numerical Methods in Engineering, 2007; Volume 69, Issue No. 5, pp. 1085-1107.
[70] Meruane, V.; Heylen, W. Structural damage assessment under varying temperature conditions, Structural Health Monitoring, 2012, Volume 11, Issue No. 3, pp. 345-357.
[71] Kuang, K. S. C.;Maalej, M.; Quek, S. T.An Application of a Plastic Optical Fiber Sensor and Genetic Algorithm for Structural Health Monitoring, Journal of Intelligent Material Systems and Structures, 2006; Volume 17, Issue No. 5, pp. 361-379.
[72] Nag,A.;Mahapatra, D. R.;Gopalakrishnan, S. Identification of delamination in composite beams using spectral estimation and a genetic algorithm, Smart Mater. Struct., 2002; Volume 11, pp. 899-908.
[73] Khatir, S.; Belaidi, I,; Khatir, T.; Hamrani, A.; Zhou, YL.; Wahab, MA. Multiple damage detection in unidirectional graphite-epoxy composite beams using particle swarm optimization and genetic algorithm, MECHANIKA, 2017; Volume 23, Issue No. 4, pp.514-521.
[74] Santos, A.; Figueiredo, E.; Silva M.; Santos R.; Sales C.; Costa, J. C. W. A. Genetic-based EM algorithm to improve the robustness of Gaussian mixture models for damage detection in bridges, Struct. Control Health Monit., 2017; Volume 24, Issue No. 3.
[75] Cha, Y.-J.; Buyukozturk, O. Structural Damage Detection Using Modal Strain Energy and Hybrid Multiobjective Optimization, Computer-Aided Civil and Infrastructure Engineering, 2015; Volume 30,pp. 347–358.
[76] Zadeh, L. A. Fuzzy sets, Information and Control, 1965; Volume 8, Issue No. 3, pp. 338-353.
[77] Hajek, P.; Paris, J.; Shepherdson, J. The liar paradox and fuzzy logic, J. Symb. Logic, 2000; Volume 65, pp. 339-346.
[78] Raju, K.; Majumdar, A.Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems, ACM Trans. Database Syst., 1988; Volume 13, Issue No. 32, pp. 129-166.
[79] Hajek, P.; Havranek, T.; Jirousek, R.Processing Uncertain Information in Expert Systems, CRC Press, Boca Raton, 1992.
[80] Juuso, E. K. Intelligent systems design with linguistic equations, 9th Workshop Fuzzy Control des GMA-FA, Dortmund, Germany, 1999; pp. 177-196.
[81] Liu, X.; Ma, L.; Mathew, J. Machinery fault diagnosis based on fuzzy measure and fuzzy integral data fusion techniques, Mechanical Systems and Signal Processing, 2009; Volume 23, Issue No. 3, pp. 690-700.
[82] Rodríguez, P. V. J.; Arkkio, A.Detection of stator winding fault in induction motor using fuzzy logic, Applied Soft Computing, 2008; Volume 8, Issue No. 2, pp. 1112-1120.
[83] Sazonov, E. S.; Klinkhachorn, P.; Gangarao, H. V. S.; Halabe, U. B. Fuzzy Logic Expert System for Automated Damage Detection from Changes in Strain Energy Mode Shapes, Nondestructive Testing and Evaluation, 2002; Volume 18, Issue No. 1, pp. 1-20.
[84] Aljoumaa, H.; Soffker, D. Condition Monitoring and Classification Approach based on Fuzzy-Filtering, Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, 2010.
[85] Reda Taha, M. M.; Lucero, J. Damage identification for structural health monitoring using fuzzy pattern recognition,Engineering Structures, 2005; Volume 27, Issue No. 12, pp. 1774-1783.
[86] Worden, K.; Delieu-Barton, J. M.An overview of intelligent fault detection in systems and structures, Structural Health Monitoring, 2004; Volume 3, pp. 85-98.
[87] Altunok, E.; Reda Taha, M. M.;Epp, D. S.; Mayes, R. L.; Baca, T. J. Damage Pattern Recognition for Structural Health Monitoring Using Fuzzy Similarity Prescription, Computer-Aided Civil and Infrastructure Engineering, 2006; Volume 21, pp. 549-560.
[88] Ettefagh, M.M.; Sadeghi, M.H.;Khanmohammadi, S.Structural Damage Detection, Using Fuzzy Classification and ARMA Parametric Modeling, Mech. & Aerospace Eng. J., 2007; Volume 3, Issue No. 2, pp. 85-98.
[89] Yu, L.; Zhu, J-H.; Yu, L-L. Structural Damage Detection in a Truss Bridge Model Using Fuzzy Clustering and Measured FRF Data Reduced by Principal Component Projection, Advances in Structural Engineering, 2013; Volume 16, Issue No. 1, pp. 207-217.
[90] Jiang, S-F.; Zhang, C-M.; Zhang, S. Two-stage structural damage detection using fuzzy neural networks and data fusion techniques, Expert Systems with Applications, 2011; Volume 38, pp. 511-519.
[91] Beena, P.; Ganguli, R. Structural damage detection using fuzzy cognitive maps and Hebbian learning, Applied Soft Computing, 2011; Volume 11, pp. 1014-1020.
[92] Liu, Z.; Zhou, Q; Chi, Q.; Zhang, Y.; Chen, Y.; Qi, S. Structural Damage Detection Based on Semi-Supervised Fuzzy C-means Clustering, The 9th International Conference on Computer Science & Education (ICCSE 2014), 2014.
[93] Irwansyah,E.; Hartati, S.; Hartono. Three-stage Fuzzy Rule-Based Model for Earthquake Non-Engineered Building House Damage Hazard Determination, JACIII, 2017; Volume 21, Issue No. 7, pp. 1298-1311.
[94] Janalipour, M.; Mohammadzadeh, A. Building Damage Detection Using Object-Based Image Analysis and ANFIS From High-Resolution Image (Case Study: BAM Earthquake, Iran), IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016; Volume 9, Issue No. 5, pp. 1937-1945.
[95] Ye, X.; Qin, Q.; Liu, M.; Wang, J.; Wang, J. Building damage detection from post-quake remote sensing image based on fuzzy reasoning, IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, 2014, pp. 529-532.
[96] Rumelhart,D. E. .; McClelland, J. L.; the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume I: Foundations, MIT Press, Cambridge, MA, 1986.
[97] Wu, X. .; Ghaboussi, J.; Garrett Jr., J.H. Use of neural networks in detection of structural damage, Computers & Structures, 1992; Volume 42, Issue No. 4, pp. 649-659.
[98] Zhao, J.; Ivan, J. N.; DeWolf, J. T. Structural Damage Detection using Artificial Neural Networks, Journal of Infrastructure Systems, 1998; Volume 4, Issue No. 3, pp. 93-101.
[99] Rhim, J.; Lee, S.W. A neural network approach for damage detection and identification of structures, Computational Mechanics, 1995; Volume 16, Issue No. 6, pp. 437-443.
[100] Pandey, P. C.; Barai, S. V. Multilayer Perceptron in Damage Detection of Bridge Structures, Computers & Structures, 1995; Volume 54, Issue No. 4, pp. 597-608.
[101] Elkordy, M. F.; Chang, K. C.; Lee, G. C. Neural networks trained by analytically simulated damage states, Journal of Computers in Civil Engineering, 1993; Volume 7, Issue No. 2, pp. 130-145.
[102] Hung, S-L.; Kao, C. Y. Structural damage detection using the optimal weights of the approximating artificial neural networks, Earthquake Engineering & Structural Dynamics, 2002; Volume 31, pp. 217-234.
[103] Wu, Z.; Xu, B.; Yokoyama, K. Decentralized Parametric Damage Detection Based on Neural Networks, Computer-Aided Civil and Infrastructure Engineering, 2002; Volume 17, pp. 175-184.
[104] Suresh, S.; Omkar, S. N.; Ganguli, R.; Mani, V. Identification of crack location and depth in a cantilever beam using a modular neural network approach, Smart Materials and Structures, 2004; Volume 13, pp. 907-915.
[105] Sahin, M.; Shenoi, R.A. Quantification and localization of damage in beam-like structures by using artificial neural networks with experimental validation, Engineering Structures, 2003; Volume 25, pp. 1785-1802.
[106] Lee, J. J.; Lee, J. W.; Yi, J. H.; Yun, C. B.; Jung, H. Y. Neural networks-based damage detection for bridges considering errors in baseline finite element models, Journal of Sound and Vibration, 2005; Volume 280, pp. 555-578.
[107] Mehrjoo, M.; Khaji, N.; Moharrami, H.; Bahreininejad, A. Damage detection of truss bridge joints using Artificial Neural Networks, Expert Systems with Applications, 2008; Volume 35, pp. 1122-1131.
[108] Kao, C. Y.; Hung, S-L. Detection of structural damage via free vibration responses generated by approximating artificial neural networks, Computers and Structures, 2003; Volume 81, pp. 2631-2644.
[109] Bakhary, N.; Hao, H.; Deeks, A. J. Structure Damage Detection Using Neural Network with Multi-Stage Substructuring, Advances in Structural Engineering, 2010; Volume 13, Issue No. 1, pp. 95-110.
[110] Yeung, W.T.; Smith, J.W. Damage detection in bridges using neural networks for pattern recognition of vibration signatures, Engineering Structures, 2005; Volume 27, pp. 685-698.
[111] 111. Nazarko, P.; Ziemiański, L. Application of artificial neural networks in the damage identification of structural elements, Computer Assisted Mechanics and Engineering Sciences, 2011; Volume 18, pp. 175-189.
[112] Haryanto, I.; Setiawan, D.; Budiyono, A. Structural Damage Detection Using Randomized Trained Neural Networks, Intelligent Unmanned Systems: Theory and Applications, 2009; Volume 192, pp. 245-255.
[113] Bakhary, N.; Hao, H.; Deeks, A. J. Damage detection using artificial neural network with consideration of uncertainties, Engineering Structures, 2007; Volume 29, pp. 2806-2815.

[114] Chang, C. C.; Chang, T. Y. P.; Xu, Y. G. Structural Damage Detection Using an Iterative Neural Network, Journal of Intelligent Material Systems and Structures, 2000; Volume 11, pp. 32-42.
[115] Chatterjee, S.; Sarkar, S.; Hore, S.; Dey, N.; Ashour, A.S.; Balas, V.E. Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings. Neural Comput. Appl., 2017; Volume 28, Issue No 8.
[116] Nia, K. R.; Mori, G. Building Damage Assessment Using Deep Learning and Ground-Level Image Data, 14th Conference on Computer and Robot Vision (CRV), 2017; pp.95-102.
[117] Kourehli, S.S. Damage Assessment in Structures Using Incomplete Modal Data and Artificial Neural Network, International Journal of Structural Stability and Dynamics, 2015; 17 pages.
[118] Abdeljaber, O.; Avci, O.; Kiranyaz, S.; Gabbouj, M.; Inman, D. J. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks, Journal of Sound and Vibration, 2017; Volume 388, pp. 154-170.
[119] Hefny, H. A.; Bahnasawi, A. A.; Abdel Wahab, A. H.; Shaheen, S. I. Logical radial basis function networks a hybrid intelligent model for function approximation, Advances in Engineering Software, 1999; Volume 30, pp. 407-417.
[120] Zheng, S-J.; Li, Z-Q.; Wang, H-T.; A genetic fuzzy radial basis function neural network for structural health monitoring of composite laminated beams, Expert Systems with Applications, 2011; Volume 38, pp. 11837-11842.
[121] Aydin, K; Kisi, O. Damage detection in structural beam elements using hybrid neuro fuzzy systems, Smart Structures and Systems, 2015; Volume 16, Issue No. 6, pp.1107-1132.
[122] Izadi, M.; Mohammadzadeh, A.; Haghighattalab, A. A New Neuro-Fuzzy Approach for Post-earthquake Road Damage Assessment Using GA and SVM Classification from QuickBird Satellite Images, Journal of the Indian Society of Remote Sensing, 2017; pp. 1-13.
[123] Janalipour, M.; Mohammadzadeh, A. A Fuzzy-GA Based Decision Making System for Detecting Damaged Buildings from High-Spatial Resolution Optical Images, Remote Sens. 2017; Volume 9, Issue No. 4.
[124] Brownjohn, J. M. W. Structural health monitoring of civil infrastructure, Phil. Trans. R. Soc. A, 2007; Volume 365, pp. 589-622.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 347-353 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Zebrafish Larvae Classification based on Decision Tree Model: A Comparative Analysis

Bayan AlSaaidah[*,1], Waleed Al-Nuaimy[2], Mohammed Rasoul Al-Hadidi[3], Iain Young[4]

[1]*Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3GJ, UK*

[2]*Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK*

[3]*Department of Computer Engineering, Al-Balqa Applied University, Al-Salt, Jordan*

[4]*Institute of Integrative Biology, University of Liverpool, Liverpool, UK*

ARTICLE INFO

Article history:
Received: 18 July, 2018
Revised: 09 August, 2018
Accepted: 21 August, 2018

Keywords:
Classification
CART Model
Feature Extraction
Zebrafish

ABSTRACT

Screening the abnormal development of the zebrafish embryos before and after being hatched for a large number of samples is always carried out manually. The manual process is presented as a tedious work and low-throughput. The single female fish produce hundreds of eggs in every single mating process, the samples of the zebrafish embryos should be studied and analyzed within a short time according to the fast response of their bodies and the ethical regulations. The limited number of the automatic screening systems for aquaculture experiments encourage researchers to find out a high-throughput screening systems with a fast prediction results according to the large number of experimental samples. This work aims to design an automatic segmentation, classification system for zebrafish eggs using two ways for feature extraction and also a classifier. Using the whole image generally with several feature vectors useful for detection process, this way does not depend on the type of the image. The second way focus on specific characteristics of the image which are the colour and the texture features relating to the system purposes. Two different ways for feature extraction integrated by the Classification And Regression Tree (CART) classifier are proposed, analysed, and qualified by comparing the two methods performance and accuracies. The experimental results for zebrafish eggs classification into three distinct classes: live egg, live embryo, dead egg show higher accuracy using the texture and colour feature extraction with an accuracy 97% without any manual intervention. The proposed system results very promising for another type of classification such as the zebrafish larva deformations.

1. Introduction

This paper is an extended work for the published paper in International Conference on Information and Communication Systems (ICICS) [1]. The proposed procedure is a part of an integrated detection, classification, counting system for zebrafish embryo malformations. After adding different chemical substances with different concentration, several deformation types appear on the larva body whether before or after being hatched. The malformations classified depending on the affected part such as the tail curvature, necrosed yolk, and the dead larva. Over the

recent years the zebrafish has become one of the most common animal models. This is due to many factors including a high degree of genetic similarity with humans, short generation times, transparent larval stages, extensively annotated genome and simple husbandry [2][3]. Zebrafish are now widely used in drug development, to measure the impact of environmental changes, of toxins and pollutants and many other applications. However, the use mammals in the biological experiment is expensive and laborious, it also led to an increasing number of

* Bayan AlSaaidah, University of Liverpool, Email: bayan@liv.ac.uk

ethical issues for toxicological research, and that have been limited in large-scale screening efforts [4].

The standard protocols, laws, and regulations for animal protection have been submitted to use zebrafish larvae in the experiment instead of using the adult and also with a specific age depending on the degree of the acuteness of the added materials, as only larvae less than five days after fertilization are exempt from this legislation [5][6]. Screening zebrafish larvae development and evaluating the effects of the chemical compounds is started from the early ages of the samples. This process is carried out manually for a large number of samples. After each mating time, the single female can produce hundreds of eggs [7] that differ depending on their hormones [8].

In [9], the author segmented the embryos from the collected images and classify the sample into live or coagulated using Bayes classifier. However, the image gathering process was carried out using camera over the microscope, this way need to be aware about several factors like illumination and the focal point for high performance capturing process. The capturing process is carried out using different focal points till reach the suitable one. The proposed classification algorithm was Bayes model using colour and shape features. The same dataset of [9] was used by [10] to classify the images into healthy and coagulated eggs. This texture features were extracted using Segmentation-based Fractal Texture Analysis (SFTA) with the rotation forest classifier.

Using video analysis for zebrafish embryos was proposed in [11] to classify if the sample is alive or dead. By detecting the zebrafish embryo heart position and analysing the intensity variation of the heart, if the cyclic motion areas are zero then the sample will be considered as dead embryo.

Despite the rapid growth in the use of zebrafish embryos as an experimental model, there is still a lack of automated classification systems according to several challenges. Screening the zebrafish eggs and larvae abnormalities within one system where the zebrafish embryos grow in a fast way and hatched within few hours. To overcome this limitation global features can be extracted regardless the type of image. The samples may slide to the edges of the petri dish this can affect the detection process. To address this problem a fixed platform is used which is a flatbed scanner with the samples container.

Another challenge is presented by unwanted objects such as the debris or food particles, as these can interfere with the detection process. To avoid this problem pre-processing operations and robust shape with colour thresholding processes are applied. The fast development of the zebrafish embryos lead to have two classes of the live embryos which have different age and different features, one of them is transparent and the other having a completed growing embryo. To address this challenge, robust and consistent different texture and colour features are extracted and used for classification purposes.

The lack of automated capture systems limit the development of a fully automated system. Using the proposed platform which is affordable and easy to use, the images are collected automatically and the biologist only need to place the dish on the scanning area.

2. Experiment and Methodology

The proposed methods aim to identify the status of zebrafish eggs using two different methods of feature extraction. The images are comprised of many features which could be extracted aautomatically or manually. These two types of feature extraction are applied and analysed for the system purposes. The collected images were gathered using a scanner collecting a large number of high resolution images (suitable for biological observation) every scan facilitating high-throughput analysis. The images were classified as: live egg, dead egg or live embryo. The scanning process is always carried out synchronously with the biological experiments. Experiments ran for up to five days post fertilization (after which the embryos become protected) during which time chemical (e.g. application of different chemicals to the holding water) or physical interventions (e.g. temperature) can be applied. The images were prepared and subdivided into three different categories for the software design step.

2.1. Data Collection

The sample images that were used in this work were collected from dish containing a hundreds of zebrafish eggs. The dish images were used in [1] for counting purposes and have been collected by the authors during several biological experiments in the Institute of Integrative Biology laboratories at the University of Liverpool in collaboration with the colleagues in this Institute. The eggs were subjected to a number of chemical substances such as: Dimethyl sulfoxide (DMSO), Alcohol, the waste nitrogenous compounds Ammonia, (Sodium) Nitrite, (Sodium) Nitrate and metals such as Copper (Sulphate) as well as antimicrobial aquarium treatments.

The images were collected using a flatbed document scanner with a high speed scanning for a petri dish of 100mm diameter. The benefit of using a scanner were manifold. The scanner has its own source of illumination and a fixed focal length, therefore, there is no need to consider the exposure or focus on the capturing process. The scanner provides the biologist hundreds of sample images in one scan where the traditional way for collecting images in the proposed systems is always carried on using camera with a microscope which is considered as a time consuming process. Figure 1 shows an example of the collected scanner images. The images that have been used in the proposed classifier were gathered from the dish images.

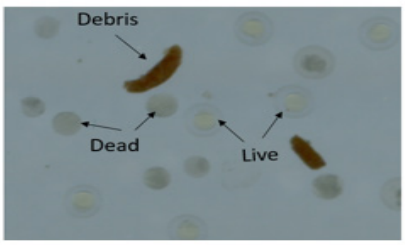


Figure 1: Part of scanner image with hundreds of eggs

Depending on the egg detection process, the samples are cropped using the egg centres and diameters. The egg detection process starts by Canny edge detection function using a threshold produced from a Sobel edge detection function. This process is followed by several morphological operations. Finally, the circles with specific range of radius are extracted from the whole image. The resulting images are collected and have been used to classify them into three classes as resulting from the biological experiments. Figure 2 shows an example from the collected images from the three classes. The live eggs between 0-40 hpf (hours post fertilization) are transparent and looks like a yellow circle. The dead eggs are white and appear dark. The last class which present a bigger embryo with more than 48hpf and before hatching.

Figure 2: Scanner images of three classes :(a) Live egg. (b) Live embryo. (c)Dead egg

2.2. Experiment

In this work, two methods of feature extraction are proposed, analysed, and assessed to develop a high performance classification system. The first method depends up on the image pixel values after splitting each colour image into multiple sub-images, while the second method deals with the sample objects and extracts the most important colour and texture features. With these two different ways the same classifier type with the same parameters is used such as the number of trees of the CART model. These two scenarios are briefly illustrated in the following two sections.

2.3. Colour and Texture Features

Traditionally, feature extraction is one of the most important steps in object recognition. The process of feature extraction uses the most valuable characteristics of the image and converts them into numeric representations. The first proposed method for our system is shown in Figure 3.

2.3.1. Image Pre-Processing

The collected images for the proposed classification system are 1211 images from the three classes live egg, live embryo, and dead egg. Several image processing operations are applied as shown in Figure 4 to segment the sample object from the whole image to make sure that the most important and useful features will be extracted where the image may contain debris and unwanted objects. These operations start with the edge detection process using Canny filter. After detection of the edges, some dilation using a disk mask with two pixels is done. To remove any unwanted objects, the largest object is segmented depending on the object areas. The resulting image is produced from the gray image multiplied by largest object binary image to keep the target object information. The processed images have been used in the next steps to extract the features using two different ways. These features are used to train the classifier how to predict and classify the three classes.

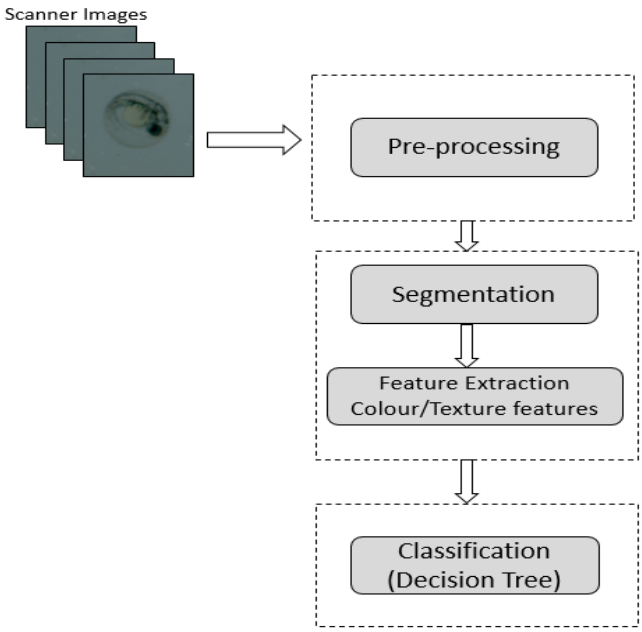


Figure 3: First scenario for classification

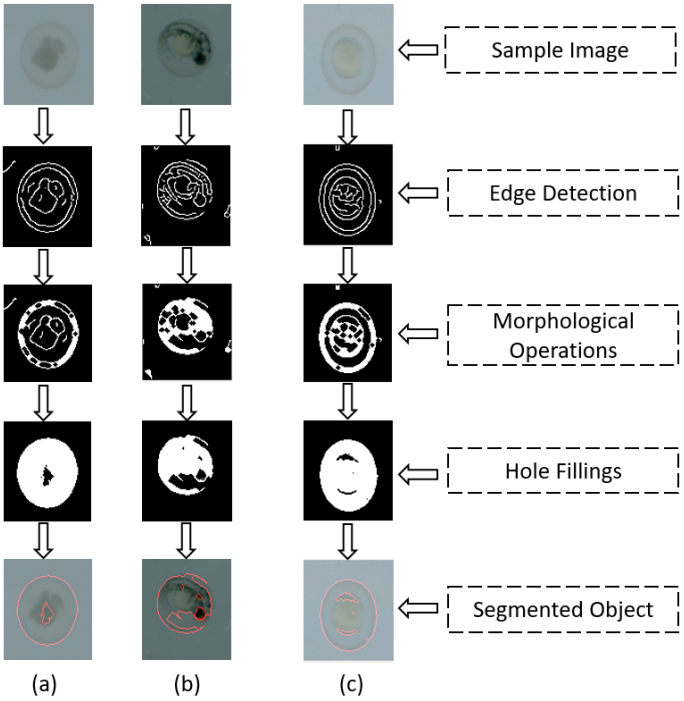


Figure 4: Object segmentation :(a)Dead egg. (b)Live embryo. (c)Live egg

2.3.2. Feature Extraction

The first attempt was done by focusing on two first order colour features of the egg according to the colour variance between the three classes. According to the colour similarity between the dead and the embryo classes, 22 texture features are

extracted from the egg images. The texture features are useful for the classification process when a wide variation of the gray levels are present in the image. Combining first and second order features aims to have robust features for classifier training step. The feature vector consists of 24 features for colour and texture image characteristics. For the image *I* of *n*x*m* in size, the mean and the standard deviation values are calculated as follows:

$$mean = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}}{mn} \quad (1)$$

$$std = \sqrt{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}(I_{ij} - mean)^2} \quad (2)$$

The first order features provide a basic information about the gray level distribution. However, the relative positions of these gray levels have not been provided by the first order features. The second order features describe and analyse if the low gray levels are together or mixed with the high gray levels. These features are calculated as proposed and illustrated in [12] [13] [14]. The second order statistics are calculated depending on a matrix $C_{\theta,d}$ (I_{p1},I_{p2}) of the relative frequencies that describes how often the two pixels (I_{p1},I_{p2}) of different or similar gray levels N_g appear as a pair in the image matrix concerning the distance *d* and the direction θ. The value of this parameter *Ng* is 8 levels.

Using the co-occurrence matrix, 22 features are extracted where the smoothness, coarseness, and the image texture information are described and quantified. The image contrast, correlation, cluster shade, cluster prominence, energy, homogeneity, entropy, and variance are measured as following:

1. Auto correlation:

$$f_1 = \sum_i \sum_j (ij)(C(i,j)) \quad (3)$$

2. Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} C(i,j) \middle| |i-j| = n \right| \quad (4)$$

3. Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij) C(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (5)$$

Where:

$$\mu_x = \sum_i i\, C_x(i)$$

$$\mu_y = \sum_j j\, C_y(j)$$

$$\sigma_x = \sum_i C_x(i) - \mu_x(i)$$

$$\sigma_y = \sum_j C_y(j) - \mu_y(j)$$

4. Cluster prominence:

$$f_4 = \sum_i \sum_j (i + j - \mu_x - \mu_y)^4\, C(i,j)$$

5. Energy:

$$f_5 = \sum_i \sum_j C(i,j)^2$$

6. Entropy:

$$f_6 = \sum_i \sum_j C(i,j) \log\left(C(i,j)\right)$$

7. Homogeneity:

$$f_7 = \sum_i \sum_j \frac{1}{1 + (i-j)^2}\, C(i,j)$$

The rest of the 15 features are correlated using Matlab functions, cluster shade [13], dissimilarity [13], homogeneity using a Matlab function, maximum probability [13], sum of squares [12], sum average [12], sum variance [12], sum entropy [12], difference variance [12], difference entropy [12], information measure of correlation1 [12], information measure of correlation2 [12], inverse difference [14], inverse difference normalized [14], inverse difference moment normalized [14].

2.3.3. Classification

The idea of the CART classifier model is presented by conditions. In this model, several questions are answered by the trees sequentially like *If-Then* condition statements. These questions depend on the extracted features from the images. Using CART model related to its efficiency and flexibility. The tree model is easy to interpret and modify according to the observed internal work. The classification consists of two main steps, training and testing. The data set is divided for training and testing processes as follows:

Table 1: Data Set Division

Class	Training Set	Testing Set	Total
Dead	322	161	483
Live (Embryo)	14	7	21
Live (Egg)	464	231	695

In the training stage the feature set (predictors) with class labels (responses) are used to train a CART classifier model. The second step is the testing step in which the classifier performance appears as an important indication of its capability. To predict the class of a new sample, the designed model follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. By repeating the first steps for preparing the image and extracting the 24 features, these features (predictors) are provided to the saved classifier model to predict the class (response).

2.4. Pixel Values Features

This approach based on the basic information of the image and without need for image pre-processing or any calculations for feature extraction. Figure 5 shows the proposed methodology.

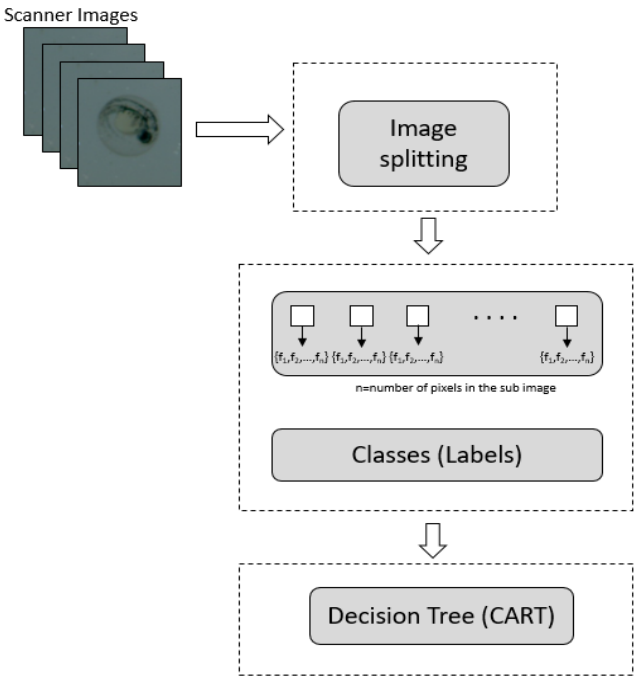


Figure 5: Second scenario of system

2.4.1. *Image Splitting*

The proposed method breaks the sample image down into several sub images. Each sub image presents a part of the whole image with a specific size (10x10) pixels from random positions and locations. The number of these sub images is chosen by trials as a sensitivity test for the process and it appears not to have a strong influence on recognition performance. However, the time complexity rises with increasing the number of extracted sub images. Each sub image is described by a feature vector containing a numerical values which are the pixel values. Using a fixed size for each sub image allows the classifier model to learn using the same size of the feature vectors for these sub images. The positivity of this process in its generalization where it can be applied on any type of images and it does not need for any image processing operations or features calculations.

2.4.2. *Classification*

Based on the ensemble approach in machine learning, a CART classifier model is built for each sub image. Each sub image is labelled according to the original image class then a CART model for each sub image is designed and provided by the feature vectors. According to the large number of features, the learning algorithm should deal with this number efficiently where this is one of the decision tree capabilities. In the training stage, the trees are grown until the model learns perfectly the training sample characteristics and can correctly classifies the training sample classes.

For the testing stage, non-labelled images are used and all the training images are no longer used here. To predict a new image, the image splitting or dividing step is repeated and the resulting vectors are provided to the saved models. Each CART model predict class depending on the sub image features. The most frequent predicted class is considered to be the final decision of the predicted class.

2.5. *Results and Discussion*

Both of the two methods of feature extraction and classification show good results. However, the texture and colour features extraction method has a higher accuracy especially for the live embryos class. This is related to the similarities between the live embryos class and the dead class. The classification algorithm is the same for the two methods where the comparison is done to compare two feature extraction ways. The first method based on the object texture and colour characteristics. The proposed work uses a CART model after extract 24 features as explained earlier. Based on this method, the classification accuracy to classify the sample image into three classes is 97%. This proposed method correctly classify 384 images out of 399 images that were used for testing. Figure 6 shows the confusion matrix for these three classes using this method.

Predicted Class

Actual Class	Dead	Embryo	Live
Dead	96%	0%	4%
Embryo	0%	100%	0%
Live	3%	0%	97%

Figure 6: Confusion matrix of the first method

Depending on the pixel values of the image is the second proposed method in which the most important features are the pixel values regardless the image type. Furthermore, this method does not need for image preparing operation or a various type of calculations for feature extraction. By tuning the number of extracted sub images parameter to different numbers, the training and testing accuracy have not been changed obviously as shown in figure.7. However, the time complexity is affected by this variations directly proportional as shown in Figure 8. In addition, using larger size of the sub images rise the system accuracy as presented in Figure 7.

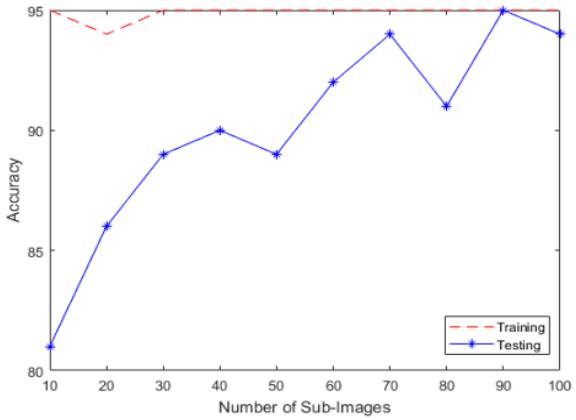


Figure 7: Classifier Performance with parameter tuning

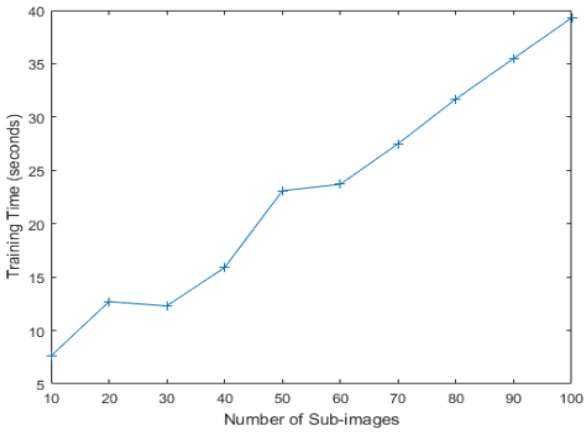


Figure 8: Classifier training time with parameter tuning

All the classification model parameters are the same in the two proposed methods. But, our analysis based on the feature extraction ways. The first one shows satisfactory results using the most important features in the image. The second way use the pixel values of the extracted sub-images to train the classifier model. This method performs well and varies according to the number of the sub-images and the size of each sub-image. The second method can be considered as a robustness and simpleness of the feature extraction calculations.

The first scenario classifier model shows a higher performance comparing with the second one, the complexity time is extremely the same. This classifier model is used to detect, segment, and classify the eggs within the whole dish image. Figure 9 shows examples of small parts of different dishes where the egg samples are detected successfully and classified correctly. The label *L* for the live organism and *D* for the dead ones.

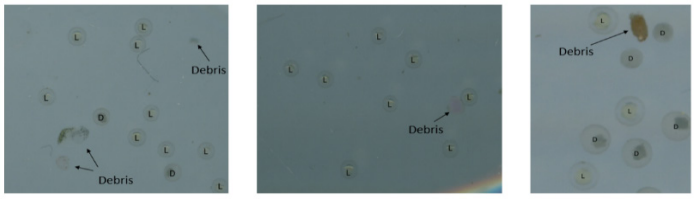


Figure 9: Classification results for the whole dish image

As shown in Figure 9, the debris which is considered as unwanted object is discarded and the target objects are detected, segmented, and classified efficiently. However, in some cases the proposed system fail in detection and classification processes as shown in Figure 10. Some samples are misclassified which can be manipulated either by cross validation for the classifier or reducing the number of extracted features by using a feature selection algorithm.

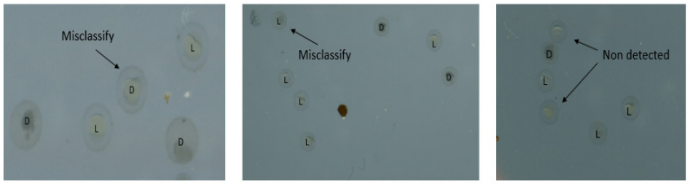


Figure 10: Drawbacks of the classifier

Using the scanner images provide us with a large number of samples of very good quality. However, these images are taken from a bigger image that contains hundreds of eggs. Partitioning process to get every sample and deal with it as a single image distort the sample images in many cases which is considered as a big challenge in this work. Table 2 shows a comparison between the scanner and the tradition way for image gathering process, camera with a microscope which considered as a time-consuming process.

Table 2: Comparison between two imaging tools

Imaging tool	Speed (image/shot)	Illumination	Focal point	Resolution
Flatbed Scanner	Hundreds of images *	Fixed by scanner	Fixed by scanner	High
Microscope & Camera	One	Important and influential (fixed by the expert)	Important and influential (fixed by the expert)	High

*depending on the number of samples, the imaging process is carried out for the whole dish.

3. Conclusion

In this paper, a novel zebrafish egg classification method is proposed using two methods for feature extraction process. Two methods were tested, analysed, and compared to each other and to the ground truth. By extracting the most important features for both colour and texture characteristics of the image, a high performance model is presented and evaluated with 97% testing accuracy to classify the sample image of the zebrafish egg into three classes depending on its status.

Using the flatbed scanner is presented as low cost effective imaging tool that save the consuming time where the one shot provide the system by hundreds of sample images. Furthermore, this tool affordable and easy to use by the biologists with a least imaging problems.

Using the pixel values of the split image parts was considered as a generic method where there is no need for applying any calculations or operations or even image modification. The results of this model is satisfactory and the parameters were tuned to a several values to reach the optimal parameter values and to see to how extent these parameters effect on the system accuracy. The performance of the first method was higher than the second one even though this method is generic and may applied on more than three classes with different characteristics.

Besides the benefits that are provided by using the flatbed scanner for data collection. The proposed system is assessed by the biologists as an effective and time saving process for their experiments. The traditional way for capturing images for the

samples is always carried out individually for hundreds of samples which is a time consuming process. In addition, the biologist need to analyse each sample to decide its status which is also considered as a time consuming process. By using the proposed system the biologist intervention is limited and the experiment time should be shorter than usual. This system present a part from a bigger classification system for detection, classification, and counting system for zebrafish embryo abnormalities using a high-throughput model for both the petri dish and the n-well plates.

4. References

[1] Al-Saaidah, B., Al-Nuaimy, W., Al-Hadidi, M., and Young, I., (2018, April). Automatic Counting System for Zebrafish Eggs using Optical Scanner. The 9th International Conference on Information and Communication Systems, 2018 9th International Conference. IEEE.

[2] Detrich, H. William, Monte Westerfield, and Leonard I. Zon, eds. The zebrafish: Disease models and chemical screens. Vol. 3. Academic Press, 2011.

[3] Yang, Lixin, Nga Yu Ho, Rüdiger Alshut, Jessica Legradi, Carsten Weiss, Markus Reischl, Ralf Mikut, Urban Liebel, Ferenc Müller, and Uwe Strähle. Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. Reproductive Toxicology 28, no. 2 (2009): 245-253.

[4] Bhusnure, O., Mane, J.M., Gholve, S.B., Thonte, S.S. and Giram, P.S., 2015. Drug target screening and its validation by zebrafish as a novel tool. Pharm. Anal. Acta, 6(426), pp.10-4172.

[5] Braunbeck T., Böttcher M., Hollert H., Kosmehl T., Lammer E., Leist E., Rudolf M. and Seitz N., Towards an Alternative for the Acute Fish LC50 Test in Chemical Assessment:The Fish Embryo Toxicity Test Goes Multi-species – an Update. ALTEX - Alternativen zu Tierexperimenten. vol.22, pp. 87–102, (2005).

[6] Al-Jubouri Q., Al-Nuaimy W., Al-Taee M., Luna J., and Sneddon L., An automatic pattern detection method for behavioral analysis of zebrafish larvae, The International Multi-Conference on Systems, Signals and Devices (SSD), Germany, pp. 301-312, (2016).

[7] Chakraborty, Chiranjib, Chi H. Hsu, Zhi Hong Wen, Chang Shing Lin, and Govindasamy Agoramoorthy. Zebrafish: a complete animal model for in vivo drug discovery and development. Current drug metabolism 10, no. 2 (2009): 116-124.

[8] Hisaoka, K.K. and Firlit, C.F., 1962. Ovarian cycle and egg production in the zebrafish, Brachydanio rerio. Copeia, pp.788-792.

[9] Alshut, Rüdiger, Jessica Legradi, Lixin Yang, Uwe Strähle, Ralf Mikut, and Markus Reischl. Robust identification of coagulated zebrafish eggs using image processing and classification techniques. In Proc. of the 19th GMA-FA 5.14" Computational Intelligence" Workshop, pp. 2-4. 2009.

[10] Tharwat, Alaa, Tarek Gaber, Mohamed Mostaf Fouad, Vaclav Snasel, and Aboul Ella Hassanien. "Towards an automated zebrafish-based toxicity test model using machine learning." Procedia Computer Science 65 (2015): 643-651.

[11] Puybareau, É., Genest, D., Barbeau, E., Léonard, M. and Talbot, H., 2017. An automated assay for the assessment of cardiac arrest in fish embryo. Computers in biology and medicine, 81, pp.32-44.

[12] Haralick, R.M. and Shanmugam, K., 1973. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, (6), pp.610-621.

[13] Soh, L.K. and Tsatsoulis, C., 1999. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Transactions on geoscience and remote sensing, 37(2), pp.780-795.

[14] Clausi, D.A., 2002. An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of remote sensing, 28(1), pp.45-62.

[15] Maree, R., Geurts, P., Piater, J. and Wehenkel, L., 2005, June. Random subwindows for robust image classification. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 34-40). IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 358-371 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Practical Approach for Extending DSMLs by Composing their Metamodels

Anas Abouzahra*, Ayoub Sabraoui, Karim Afdel

Laboratory of Computer Systems and Vision LabSIV, Ibn Zohr University, Agadir, Morocco

ARTICLE INFO

Article history:
Received: 27 August, 2018
Accepted: 21 November, 2018
Online: 05 December, 2018

Keywords:
Model Driven Engineering
Domain Specific Modeling
Domain Specific Languages
Model Composition
Software Reuse
Code Generation
Experimental Software Engineering

ABSTRACT

Domain specific modeling (DSM) has become popular in the software development field during these last years. It allows to design an application using a domain specific modeling language (DSML) and to generate an end-solution software product directly from models. However providing a new DSML is a complex and costly job. This can be reduced by the reuse of existing DSMLs to compose new ones trough a metamodel composition approach. This paper provides a composition rules based code generator facility for extending DSMLs. In doing so, it proposes three rules to compose DSMLs by composing there metamodels: reference rule, specialization rule and fusion rule. The results of an exploratory case study on using these rules are depicted. In addition a proof of concept of the code generator facility which generates the necessary infrastructure to quickly build new DSMLs is implemented and applied to the case study. The benefits of our approach are measured relying on three indicators: the reduced development time, the reused software components and the gain on learnability.

1. Introduction

This paper is an extension of work originally presented in 2017 European Conference on Electrical Engineering and Computer Science (EECS) [1]. The motivation of this paper is to improve the state of the art of quick development of new DSMLs based on existing ones.

Software composition is a fundamental mean for the evolution of complex software systems [2]. While initial approaches were simply focused on textual composition, more efficient approaches take into account syntax and semantics of the software. There was a tendency over the last twenty years towards operation based composition because of its increased expressiveness. In this direction, Model Driven Engineering (MDE) [3,4] was concerned about improving model composition approaches. From early, the researchers have realized that the application of MDE to complex systems will undoubtedly go through the development of smart and agile model composition techniques [5–9].

A use that takes advantage of model composition is to speed up the implementation of new Domain Specific Modeling Languages (DSML). Designing DSMLS is a not an easy job and generally consuming time [10]. This operation can be simplified by reusing existing DSMLs, composing their metamodels, to get new and larger ones [11]. In fact, the definition of a DSML is based on a metamodel and often provides supporting tools as graphical editors to create and handle models. Therefore, it would be judicious to define the reuse of artifacts at the abstract level; i.e. at metamodels level, then to deduce the projection of this composition at the underneath levels; i.e. the supporting tools.

In the previous work [1], composing metamodels of DSMLs was studied and consequences on their graphical editors were investigated in order to provide a composition of metamodels based approach to extend DSMLs. This work goes further and presents an exploratory study that aims to evaluate the DSML composition approach exposed in [1]. It implements a proof of concept of this approach by developing a code generator facility to make composing graphical editors of DSMLs easier. This prototype provides an automatic code generator which starts from a composition of metamodels of DSMLs, described using composition rules, and generates a layer of code allowing a rapid composition of a new DSML. The gain is measured in terms of development time that can be estimated via the percentage of the generated code. Then, in terms of reused components that can be estimated via the percentage of reused code. As well as in terms of learnability, that can be estimated via the part of kept features and interfaces. These three parameters will be the indicators of evaluation and performance to assess the contribution of this work.

The paper is organized as follows: In Section 2, a series of related works is cited. In Section 3 the problem is stated and the followed methodology is explained. In Section 4 the exploratory

* Anas Abouzahra, Email: abouzahra.anas@gmail.com

study is developed. In Section 5 results are discussed. Finally the Section 6 concludes the paper.

2. Related Works

In this Section a selection of works addressing the composition, extension and reuse of DSMLs in an MDE context is exposed with a brief summary of their features. Moreover, approaches coming from Aspect Oriented Programming (AOP) [12] and Language-oriented programming (LOP) [13] research fields are presented. That is because they have inspired relevant methods for reusing and extending DSMLs. Other approaches that use software product line (SPL) [14] techniques exist [15] but they are minor.

2.1. In MDE

MDE addressed the problem of extending DSMLs by a composition operation. However it varies according to the meaning that each work gives to the composition. This can be a merge operation between models conforming to the same metamodel. As it can be a fusion operation between completely heterogeneous models. It can also be just a resolution of differences between different versions of the same model in order to resolve existing conflicts. Besides, the composition operation can be automatically generated based on mapping calculations or completely custom based on a weaving definition. Furthermore, other approaches provide complementary operations such as checking the consistency of the composition.

Epsilon EML [16] is an Eclipse project which provided a platform for developing substantial and interoperable operations on DSLs among which there is a model composition operation (merging) [17] provided through the Epsilon Merging Language (EML) language [18,19]. EML is applied to compose a number of potentially heterogeneous models. The composition operation is achieved through four steps: comparison, conformance verification, composition and reconciliation. EML was the first language accommodate for model merging and made the case for non-trivial merging of heterogeneous models. However it turned out that it is too verbose for merging homogeneous models. Although, EML is still maintained with significant evolution of its syntax, semantics, capabilities and its underlying platform

AMW [20] is an Eclipse project which proposed a model composition solution (weaving) in parallel of a higher level transformation. In AMW, megamodeling has been introduced to tackle advanced metamodel management, where often the relationships between the metamodels can be considered as composition links. Model weaving has been often used as a solution to compose different DSLs, where the composition is not only the simple gathering of concepts coming from different metamodels, but might also include advanced semantic operators. Unfortunately, and despite all the interest in this tool and its various applications, especially for the traceability of model transformation, the associated eclipse project has been archived. It has not been maintained by the community and no longer by an industrial.

MOMENT [21,22] is a project which aimed to provide a model management platform that furnishes generic operators to handle metamodels described using the Eclipse Modeling Framework (EMF) [23]. In this context, Boronat et al. developed a practical approach for generic model merging. It provides an automate merge operator to merge DSLs artifacts with support for conflict resolution and traceability [24] relying on the QVT Relations language [25]. This work was applied, and specially proven, to class diagrams integration [26].

Melange [27,28] is a project which treated the modularity and the reuse of DSLs and brought a meta-language for implementing DSLs by composing and specializing existing DSL units. It specifies operators for language assembly, for language extensions and language restrictions. Almost introduced operators by are meant to reuse either the semantics or the metamodel as is, in addition of merging code. Except the inheritance operator which is able to modify the initial definitions in the new DSL. Nevertheless, it is not clearly explained how the extended metamodel modify the original one and which concepts can be overridden.

MetaEdit [29,30] is a graphical workbench which provided a language for creating DSMLs. It introduced the concept of joint/linked modeling constructs to reuse DSLs with code generation facility. In MetaEdit+, the code generation is obtained by the use of a template based on the target language. Consequently, it limits the modularity scope of the generation to the modularity capabilities of the target language. Nevertheless, in MetaEdit+ each created DSL is an addition to, or an extension of, the language workbench itself [31]. This extends the capabilities of reuse to DSLs that are already defined in the workbench.

Other works have treated the problem of model composition and reuse from different angles. Indeed Berg et al. in [32] propose an operational semantics based approach for composing and reusing metamodels and models, by including their operational semantics. Composition is performed relaying on a reusable template that permit customizing the metamodel meta-concepts as part of the composition operation. It uses a placeholder mechanism where given meta-concepts of a given metamodel are reused in another metamodel [33]. Schmidt et al. in [34] treated the problem of model composition from a collaborative modeling point of view. They proposed an approach to ease the merging of complex models that are collaboratively developed in teams. This approach aims to furnish collaborative development capabilities in much the same quality as it is provided by version control software or text document merging tools. A recent work in [35] contributed to the same purpose. More, it focused not only on conflicts but also on arbitrary syntactic and semantic consistency issues. Coherent artifacts are merged automatically and only conflicting artifacts are presented to the designer's attention, along with a systematic suggestion of resolution. Otherwise, some works focuses on providing complementary operations to model composition such as checking its consistency. In this direction, Zhang et al. [36] implemented WMCF which is a models composition framework relying on the Alloy language [37,38]. It furnishes a model weaving capability with consistency checking of the resulting composition provided by the Alloy Analyzer.

Besides, generating graphical editors from an abstract definition of a DSL has been addressed by many works. Notably, the *EMF Edit*, the *Graphical Modeling Framework* (GMF) [39,40] and the *Generic Modeling Environment* (GME) [41,42] had brought important contributions. They are mature frameworks based on MDE concepts and furnish tools for defining grammars and generating code for graphical editors.

GMF provides a set of capabilities and runtime infrastructures for generating graphical editors for DSMLs based on their metamodel definitions. Where GME allow decorating a metamodel of a DSML with entities called views. This gathers concepts that will be used to implement models, links between

those concepts and how the concepts will be organized and displayed by the graphical editor. Nevertheless, these frameworks, even if they make the generation of graphical editors of DSLMs much easier, they did not elaborated proper features to support the composition and reuse of DSMLs.

2.2. In AOP

The MDE was much inspired by AOP to deal with the problem of large models for complex systems [43]. The AOP preconizes to design a system by separating the model into different morsels. Each corresponds to a different aspect of the system. This decomposition permits to deal with properties on each aspect before considering the model in its overall. This way we decrease the analysis complexity [44]. However, this requires being able to integrate the morsels of a model with each other's. Thus, AOP has addressed the problem of composition and reuse of models [45].

Hovsepyan et al. [46] elaborated an asymmetric approach to compose artefacts of different DSMLs using an application base model described with UML. This approach was driven by an AOP methodology and was implemented using MDE tools. It introduced the concept of a concern interface which plays the role of a common language between a specific concern and the application base. The composition is then achieved by defining explicitly the syntactic and the semantic relationships between artifacts coming from different concerns.

LARA [47] is a DSL inspired by AOP concepts which brought a novel method for mapping applications to heterogeneous high performance embedded systems. It allows to generate an intermediate aspect representation from a configuration based on different junction points, action models and attributes. This is then given to be processed by the weavers. Pinto et al. [48] has improved LARA by furnishing well-defined library interfaces with concrete implementations for each supported target language. This work contributes to make LARA aspects more concise and improve their reuse. Moreover, it involve to substantial reductions of job effort when developing weavers for new languages.

MATA [49] is an AOP tool built on the top of IBM Rational Software Modeler. It uses model transformation to define and perform composition operations on aspects of a model. The particularity of MATA is that, even if it is inspired by AOP, it did not deal with specific join points. In fact, any model artifact can be considered a join point, and composition is implemented as a special case of model transformation. In addition, critical pair analysis is automatically applied in order to find structural correspondences between various aspects of models. MATA was intended to be a generic approach but it was above all proven on UML models (class diagrams, sequence diagrams and state diagrams) [50].

2.3. In LOP

The LOP field is rich in approaches that ease the design and the reuse of DSLMs. Ones of the most presumably technologies to perform it are the projectional language workbenches. In fact, they provide relevant approaches for extending a DSL and often furnish tools to project it on concrete spaces.

TouchRAM [51] provided a rich client tool for flexible software modelling. It enable at developing reusable and scalable design model through a large registry of design basic design models. It takes advantage from model interfaces and aspect oriented model weaving. The conception of a new design model can be obtained by the composition of available design models in the registry. This work has been improved with TouchCORE [52] which furnish new features for model visualization, model editing model assessment and composition traceability.

MPS [53] provided capabilities to define a DSL trough many aspects: abstract aspects (metamodel), sematic aspects (constraints), concrete syntaxes aspects (graphical editors), generators aspects (model transformations) and many others (e.g. behavior, type system, data flow or intentions) [54]. MPS furnish two ways to reuse DSLs: the reference and the extension mechanisms. The reference consists to use concepts from a given DSL into another one. The extension allows extending a DSL from another one by creating new concepts that inherit all the properties and behavior of their parents [55].

MetaMod [56] is based on a metametamodel that provides metatools to ease the creation and the reuse of DSLs. Convicted that most of simple DSLs do not require more advanced modularity, MetaMod defines the modularity at the value model level provided by the metametamodel itself [57]. Furthermore, having the same modularity mechanisms in many DSLs lead to have robust DSLs, because easier to verify, more fit and easier to learn as well. In addition, this facilitates the reuse of DSLs. However, it limits capabilities of the DSLs if more advanced modularity mechanisms are needed.

Cedalion [58,59] is built on top of Prolog. It provides features for DSLs building with projectional editor trough the description of model aspects such as semantics, structure, projection, and type system definitions from other language workbenches. Cedalion proposes a DSL reuse mechanism. However, because of the close link between the structure of a language and its other aspects, this makes the reuse difficult in Cedalion. In fact, all language aspects of a DSL need to be reused. Nevertheless, extending a DSL with only additional concepts is thus effortless [60].

Spoofax [61] is a language workbench dedicated to design textual DSLs. The platform provides features for code generation, parsers, type checkers, compilers, interpreters, and other tools from language definitions. Spoofax furnishes an API for programmatically composing abstract and concrete syntax of a language. Within Spoofax, the management of modularity can be managed directly in target generated language [62].

Xtext [63] is a textual language workbench based on EMF. It provides tools to define textual DSLs. It furnishes a DSL reuse and extending mechanisms. Reuse permits to cross reference concepts between DSLs. Where, extending allows to a DSL to inherit from another one and to override its grammar rules. However, it allows only to completely overriding them. In addition, this extending mechanism limits a DSL to only extend one other. Regarding the dynamic semantics of the DSL, it can be implemented using other languages such as Xtend [64].

Monticore [65,66] is a textual language workbench. It provides modularity mechanisms that enable the compositional development of textual DSLs and their supporting tools. Especially, inheritance and embedding mechanisms are proposed. Inheritance allow to extend a language where embedding allow to compose different language fragments. Moreover, a special DSL is proposed for the definition of compositional links between languages.

Other works implemented approaches to provide DSMLs composition and reuse capabilities. Pedro et al. [67] have

contributed with an automatic projection approach from metamodels composition patterns into graphical syntax. In [68] they go further more with a definition of operators to compose DSMLs with a proposal for automatic mapping to graphical syntax. Meyers et al. [69,70] proposes a template based technique for the modular definition and composition of DSMLs, including their abstract syntax, semantics, and concrete syntax (relying on metaDepth [71]).

3. Problem Statement & Methodology

Software engineering is essentially involved in providing textual or graphical languages to describe and set out artefacts of a system; their structures, behaviors and interactions. DSMLs provide capabilities to achieve this and allow designers to handle these artefacts as models. Models are intended to be used by tools. Thus, it should be defined a formal description of their concepts. This well-established set of concepts is called a metamodel. This is the principle of DSMLs design. Accordingly, composing DSMLs is primarily a composition of their metamodels.

The composition of metamodels is special issue of a larger problem in MDE, model composition. The composition of the models is a topic of research in continual but very slow evolution in the MDE. This is partly due to the miss of inspiration of patterns from programming languages [72]. It also never has been the subject to standardization like model transformation.

We can define a composition of models simply with a composition operator ⊗ which is a function producing a composed model C by using artifacts of two input models A and B:

$$\otimes : A \times B = C \quad (1)$$

However, model composition can scope various meaning and reach at least three dimensions: abstract syntax, concrete syntax, and semantics [73]. As a DSML is a modeling language we can take inspiration from the composition operation as it is defined in modeling languages; an association of sub languages into one integral language. Where sub artifacts are handled in their original languages and the composed artifact acquire its semantics and its syntax from the composition [74]. In addition, we can draw inspiration from the composition operation as it is defined in programming languages. There is a frank conjunction between the semantic unit (i.e. class) that has a specific interface and the syntactic unit (i.e. file) that is the encapsulation of the implementation [72]. When semantic units are composed, logically de facto the language tool composes the syntactic units. Therefore, a successful approach of composition of DSMLs must deal with the three composition dimensions and maintain the link between abstract syntax (metamodels), concrete syntax and semantics.

In the respect of the aforementioned, this work is an exploratory study whose purpose is to explore means of composing DSMLs by composing their metamodels and studies the projection on their associated graphical editors. Indeed, an appropriate reuse of their syntaxes and graphical editors can be performed. Defining the way that metamodels will be composed implies the way that syntaxes can be merged and editors can be reused. Furthermore, it explores how to automate the composition of graphical editors of the composed DSMLs by implementing a prototype of a composition rules based code generator facility. For that the proposed exploratory study is segmented into five steps as described in Figure 1.

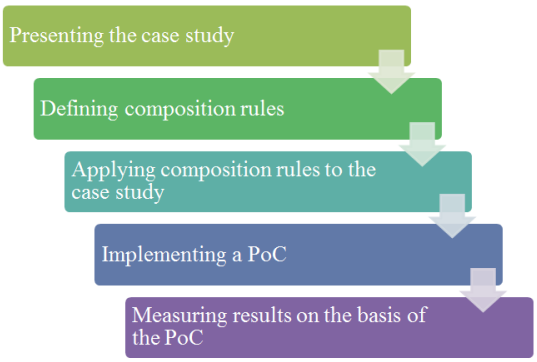


Figure 1. The process of the exploratory study.

A case study is first described. Then the rules for composition of metamodels are defined; on the basis of which DSMLs artifacts can be reused. Later, each is applied to a use case from the case study in order to illustrate it. Next, to prove the concept of this work, an implemented prototype of code generator facility for extending new DSMLs based on the aforementioned composition rules is presented. Finally, three parameters as indicators of evaluation and performance are used to assess the contribution of this study:

- The gain in terms of saved development time that can be estimated via the percentage of the generated code.
- The gain in terms of reused components that can be estimated via the percentage of reused code.
- The gain in terms of learnability, that can be estimated via the part of kept features and interfaces.

These three parameters are measured and discussed in the Section 5.

4. Exploratory Study

4.1. Case Study

Figure 2 represents excerpts of three simple metamodels representing small DSMLs. Each metamodel relies on a different concept.

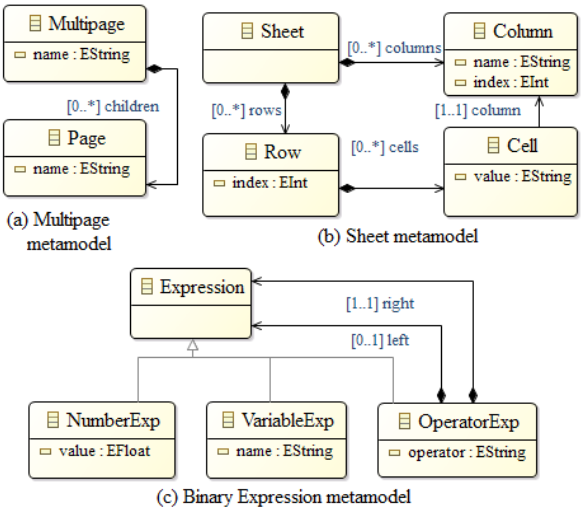


Figure 2. Multipage, Sheet and Expression metamodels.

The first metamodel is the Multipage metamodel ($MM_{Multipage}$). It can be used to describe a multiple page structure. It allows multiple pages to be contained under a single parent page. According to (a), an instance of the metaclass *MultiPage* may contain one or more children instances of the metaclass *Page*. The second metamodel is the Sheet metamodel (MM_{Sheet}). It can be used to describe a sheet containing a two dimension table. According to (b), an instance of the metaclass *Sheet* contains a collection of instances of the metaclass *Row*. Each of which containing a collection of instances of the metaclass *Cell*. In addition, the instance of *Sheet* defines instances of the metaclass *Column*. Each defined instance of *Cell* is related to one of them. The third metamodel is the Binary Expression metamodel (MM_{Exp}). It can be used to describe a binary expression tree which can contain numbers, variables, and unary or binary operators. According to (c), an instance of the metaclass *Expression* is a tree of nodes which can be instances of three metaclasses: *OperatorExp*, *IntegerExp* and *VariableExp*. The *OperatorExp* instances are contained in the internal nodes of the tree, where instances of *IntegerExp* and *VariableExp* are contained in the leaf nodes. Withal, an OperatorExp node may have two children nodes for binary operators (left and right expressions), or one child node for unary operators (only right expression). Each of these DSMLs metamodels relies on a graphical or textual syntax that allows expressing conforming models. Therewith, they are supported by graphical interfaces: editors. Figure 3 shows screen shots of the editors. The graphical syntax of the $MM_{Multipage}$ DSML (a) expresses a *MultiPage* instance as a multiple tabs window. The children instances of *Page* are embedded as a sequence of tabs in the parent *MultiPage* instance. The associated editor has buttons to add new tabs. The graphical syntax of the MM_{Sheet} DSML (b) expresses an instance of *Sheet* as a two dimensions table with indexed instances of *Row* and named instances of *Column*. Instances of *Cell* are represented by the boxes of the table with their *values* inside. The associated editor permits to extend or reduce the table using a hold and move button. This way, the editor allows creating new instances of *Row* and *Column* or deleting existing ones. *Indexes* of *Row* instances are represented at the left side of the table. Editing *names* of instances of *Column* and *values* of instances of *Cell* is done via the textual edition of the related boxes. The textual syntax of the MM_{Exp} DSML (c) expresses an instance of *Expression* using a mathematical grammar where parentheses represent internal nodes. The associated editor is a textual file editor with syntax highlighting.

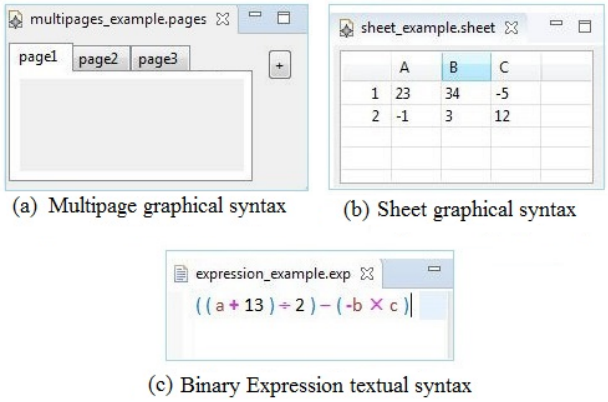


Figure 3. Multipage, Sheet and Expression concrete syntaxes.

This case study is used later to create step by step a new DSML that meets the following requirements reusing DSMLs (a), (b) and (c):

- *RQ1*. A sheet cell must be able to contain a binary expression.
- *RQ2*. A binary expression defined inside a cell must be able to reference the value of another cell of the sheet's table.
- *RQ3*. A multiple page must be able to be composed of multiple sheet tabs.

In the following Subsections three rules for composing metamodels are defined. Next, they are illustrated relying on the above requirements. Each time a requirement is fulfilled it uses an application of a defined rule. Besides, an investigation of the reuse of the syntax and graphical artifacts of original DSMLs is realized. It aims to obtain an extended DSMLs based on the performed composition of metamodels expressed by means of the proposed composition rules.

4.2. Composition Rules

To describe a composition rule, the following formalism is used :

$$\otimes\text{Rule: } MM_A \times MM_B \text{ (}arguments \ldots\text{)} = MM_C \quad (2)$$

Where;

- MM_A and MM_B are metamodels to be composed.
- ⊗Rule is the composition rule.
- MM_C is the composed metamodel.

Reference Rule

A reference rule allows the establishment of discrete connections between instances of a model, conforming to MM_C, defined by concepts coming from MM_A and MM_B. It defines an oriented binary association in MM_C from a metaclass MT_1 of MM_A toward a metaclass MT_2 of MM_B. It is used to connect one instance of MT_1 to many instances of MT_2. It could be a simple link, an aggregation or a composition. It must specify *multiplicity* to mean how many instances of MT_2 can be referenced from an instance of MT_1. The $\otimes Ref$ rule can be defined as follows :

$$\otimes\text{Ref: } MM_A \times MM_B \text{ (}MT_1, MT_2, [m_1, m_2], c_1, c_2\text{)} = MM_C \quad (3)$$

- $[m_1, m2]$ are integers to express multiplicity with a minimum value m_1 and a maximum value m_2.
- c_1 is a Boolean value to mean whether the reference expresses a containment association (i.e. aggregation).
- c_2 is a Boolean value to mean whether the reference expresses a container association (i.e. composition).
- MM_C is the composed metamodel.

Specialization Rule

The specialization rule permits to compose metamodels with an inheritance concept similar to the concept of specialization in object oriented programming. It allows to a metaclass MT_1 from a metamodel MM_A to acquire all the properties and behaviors of another metaclass MT_2 from a metamodel MM_B. Thus, attributes,

associations, or methods can be reused. The ⊗*Spe* rule can be defined as follows :

$$\otimes\text{Spe: } MM_A \times MM_B\,(MT_1, MT_2) = MM_C \quad (4)$$

Fusion Rule

The fusion rule is used to bind metaclasses coming from different metamodels in order to fusion them in the composed metamodel. It allows a metaclass MT_1 from a metamodel MM_A and a metaclass MT_2 from a metamodel MM_B to form a new hybrid metaclass MT_3 in the composed metamodel through over a customized binding. The binding defines the matching between the properties of metaclasses MT_1 and MT_2: attributes references and methods. In addition it specifies those to keep and those to delete. The ⊗*Fus* rule can be defined as follows:

$$\otimes\text{Fus: } MM_A \times MM_B\,(MT_1, MT_2, \{bindings\}) = MM_C \quad (5)$$

4.3. Rules Application

A ⊗*Ref* rule can be applied to fulfill the requirement RQ1. Considering that MT_1 is the *Cell* metaclass as defined in the MM_{Sheet} metamodel and MT_2 is the *Expression* metaclass as defined in the MM_{Exp} metamodel. The ⊗*Ref* rule can be applied as follows:

$$\otimes\text{Ref}_{\text{expression}}: \quad MM_{Sheet} \times MM_{Exp}\,(Cell, Expression, [0, 1], true, true) = MM_{C1} \quad (6)$$

By applying the $\otimes Ref_{expression}$ rule, the composed metamodel MM_{C1} is obtained. MM_{C1} is shown in Figure 4, where the $\otimes Ref_{expression}$ rule is represented with the bold line starting with a lozenge. The designed composition allows an instance of *Cell* to contain an instance of *Expression*. The achieved composition is projected in order to create a new extended graphical editor for the new DSML (d). It is based on the MM_{C1} metamodel and reuses the graphical artifacts of DSMLs (b) and (c). A mock-up of the extended editor of (d) is shown in Figure 4 where the textual editor of the DSML (c) is included in the top of the editor of the DSML (b). According to (d), a sheet's cell is able to contain the value of a binary expression. Therefore, the cell's value is the computed value of an expression which can be edited in the top textual editor.

Similarly, another ⊗*Ref* rule can be applied to fulfill the requirement RQ2. However, this time the ⊗*Ref* rule has to define a simple link reference, from the metaclass *VariableExp* toward the metaclass *Cell*. The ⊗*Ref* rule can be applied as follows :

$$\otimes\text{Ref}_{\text{refersTo}}: \quad MM_{C1} \times MM_{C1}\,(VariableExp, Cell, [0, 1], false, false) = MM_{C2} \quad (7)$$

By applying the $\otimes Ref_{refersTo}$ rule, the composed metamodel MM_{C2} is obtained. MM_{C2} is shown in Figure 4, where the $\otimes Ref_{refersTo}$ rule is represented with the bold arrow. The designed composition allows the definition of cross references between cells expression. In this way, an instance of *Expression* contained inside an instance of *Cell* can reference the value of another instance of *Cell* present in the table. Explained otherwise, an instance of *Expression* which is structured as a tree can include in its nodes a reference to an instance of *Cell* through an instance of *VariableExp*. It is important to observe that this design adapts the *Cell*'s instances to behave as *Expression*'s instances. It is worth mentioning that such pattern, applied with the ⊗*Ref* rule can be useful to solve situations where it is need to adapt concepts of different metamodels when composing them. The achieved composition is projected in order to create a new extended graphical editor for the new DSML (e). It is based on the MM_{C2} metamodel and reuses the graphical artifacts of the DSML (d). A mock-up of the extended editor of (e) is shown in Figure 4. According to (e), a binary expression defined inside a cell must be able to reference the value of another cell of the sheet's table. Therefore, the cell's value is the computed value of an expression which can use the computed values of expressions defined elsewhere in the sheet.

A ⊗*Spe* rule can be applied to fulfill the requirement RQ3. Considering that MT_1 is the *Sheet* metaclass as defined in the MM_{C2} metamodel and MT_2 is the *Page* metaclass as defined in the $MM_{Multipage}$ metamodel. The ⊗*Spe* rule can be applied as follows:

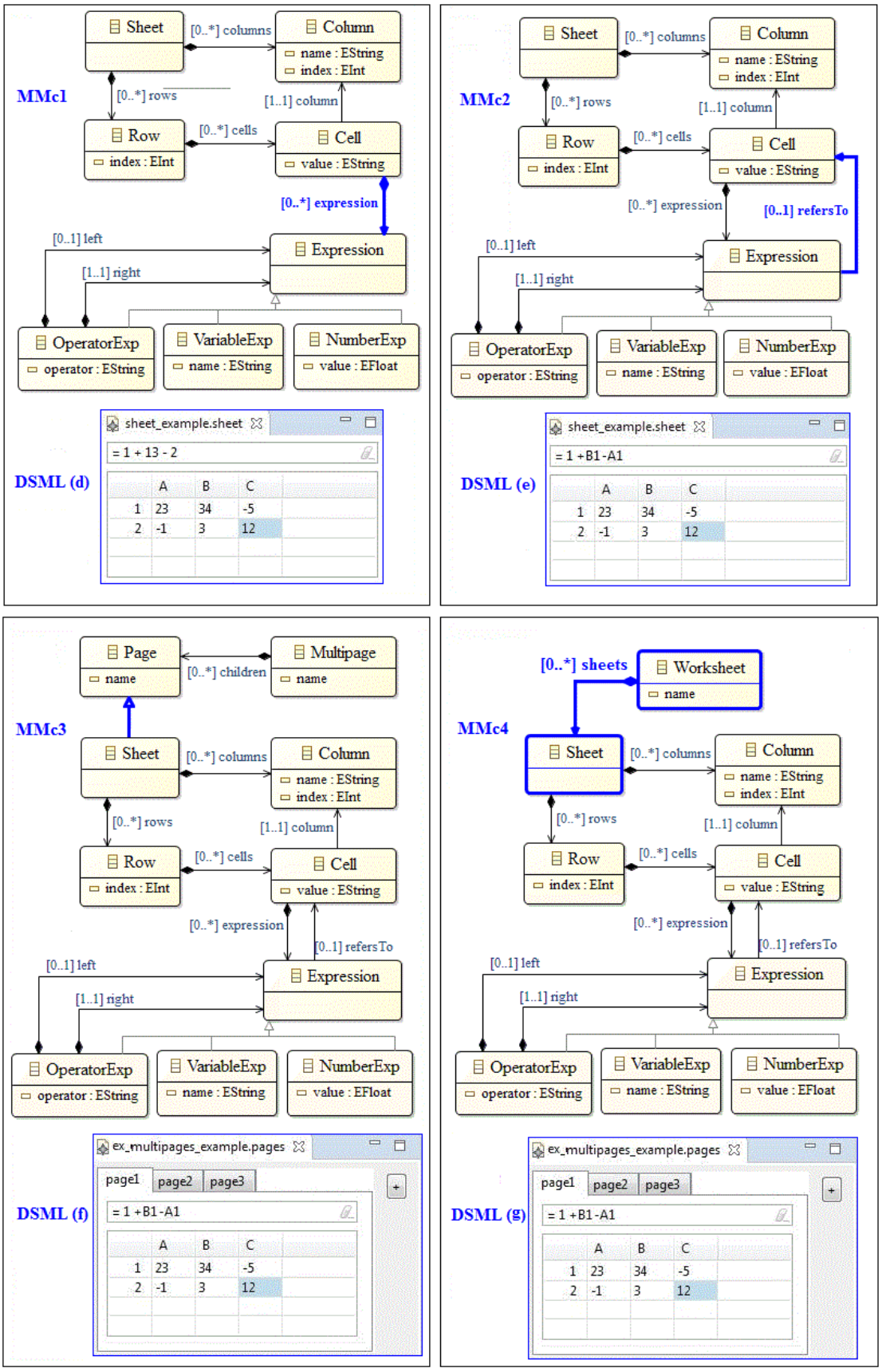
$$\otimes\text{Spe}_{\text{page}}: MMc2 \times MM_{Multipahge}\,(Sheet, Page) = MMc3 \quad (8)$$

By applying the $\otimes Spe_{page}$ rule, the composed metamodel MM_{C3} is obtained. MM_{C3} is shown in Figure 4, where the $\otimes Spe_{page}$ rule is represented with the bold arrow. The designed composition allows a sheet to be a kind of page and then to be a candidate to be a tab of the multiple page. In this way an instance of Multipage can contain instances of Sheet. Therefore, a multiple page is able to be composed of multiple sheet tabs. The achieved composition is projected in order to create a new extended graphical editor for the new DSML (f). It is based on the MM_{C3} metamodel and reuses the graphical artifacts of the DSML (f). A mock-up of the extended editor of (f) is shown in Figure 4. According to (f), it is possible to create multiple tabs of sheets using the means of the multipage editor; i.e. the button that creates pages. The graphical interface of a sheet is then embedded in the graphical container provided for a page and behaves autonomously. However a question may arise about the semantic and utility of the metaclass *Page* in MM_{C3}. The answer depends on the understanding of the requirement RQ3. If it requires obtaining a multiple pages editor composed "only" of sheet tabs. It is probably cleaner to merge metaclasses *Page* and *Sheet*. Moreover, it would be clearer to give a new semantic to the metaclass *Multipage* to indicate that it is a multiple sheets tabs. This leads to define a new rule: the fusion rule.

A ⊗*Fus* rule can fulfill the requirement RQ3 in case it requires obtaining a multiple pages editor composed only of sheet tabs. Considering that MT_1 is the *Sheet* metaclass as defined in the MM_{C3} metamodel and MT_2 is the *Page* metaclass as defined in the $MM_{Multipage}$ metamodel. The ⊗*Fus* rule can be applied as follows:

$$\otimes\text{Fus}_{\text{sheet}}: \quad MMc2 \times MM_{Multipage}(Sheet, Page, \{Sheet.name, Page.name\}) = MMc4 \quad (9)$$

By applying the $\otimes Fus_{sheet}$ rule, the composed metamodel MM_{C4} is obtained. MM_{C4} is shown in Figure 4, where the $\otimes Fus_{sheet}$ rule had the apparent effect to superimpose the metaclasses *Page* and *Sheet* in one metaclass. Additionally, the metaclass *Multipage* can be renamed to *Worksheet* in order to give a better representative name to the container of sheet tabs. The achieved composition projected in order to create the extended graphical editor for the



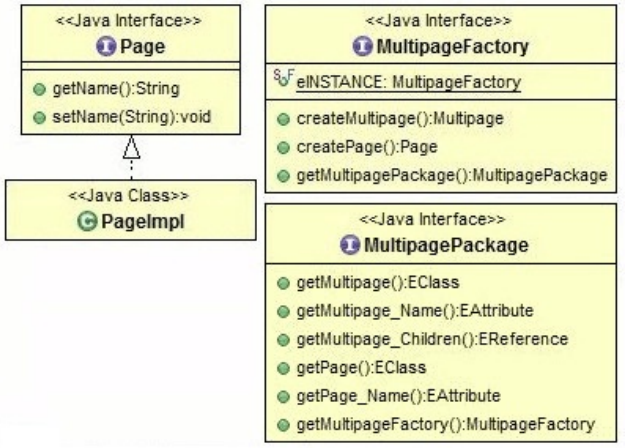
The applications of composition rules on the case study.

new DSML. It is based on the MM_{C4} and leads to the graphical editor of the DSML (g). It very close to the DSML (f) except that the content of the multiple sheets (worksheet) can only be sheet tabs. According to this new DSML, an instance of *Worksheet* can be composed, and only composed, of multiple instances of *Sheet*.

4.4. Proof of Concept

In order to validate the approach exposed in this paper and going further than the theoretical exposition, a Proof of Concept (PoC) was achieved. For this purpose, a prototype of code generator facility based on the aforementioned composition rules was implemented. Then it was applied to our case study. However, before doing so, it is necessary to implement the DSMLs (a) and (b) used in the case study. Next, the metamodels of these DSMLs are composed using one of the rules previously defined. Finally the implemented prototype is used to project the composition onto the graphical editors of DSMLs.

EMF is used to implement the PoC. EMF allows the generation of class architecture that represents metamodel concepts. EMF does not only generate Java classes, but also an associated infrastructure. Thus, one benefits from the persistence of the model in XML Metadata Interchange (XMI) [75] format, but also from a set of tools to handle the model completely independent from the objects it contains. This infrastructure makes it possible to build higher level tools for processing models created with EMF. Within this framework, one of the functionalities is notably the visualization and the edition of models thanks to the EMF Edit framework. Using the capabilities of EMF, the prototype of the PoC aims to generate an overlay layer of code following the EMF code generation. It must provide the necessary infrastructure to make quick building of new DSMLs based on the composition of their metamodels possible.




```
/**
 * Example of generated code usage
 */
// Factory to create instances of model objects
Page page = MultipageFactory.eINSTANCE.createPage();
// Package class to provide access to meta data
EAttribute name_att = MultipagePackage.eINSTANCE
                                      .getPage_Name();
// Set page name using reflective method
page.eSet(name_att, "page 1");
// or direct method
page.setName( "pag 1");
```

Figure 5. A class diagram representing EMF Impl generated classes for the metaclass Page and an example of code usage.

The first step is obviously to implement metamodels. The PoC use case relies only on the two metamodels *Multipage* and *Sheet*. The EMF essential MOF [76] implementation (Ecore) is used to describe metamodels. As aforementioned, EMF provides a code generator facility to generate Java code from a metamodel described in Ecore. It generates, inter-alia, two based packages: One for model implementation (EMF Impl) and another one for graphical user interface editing (EMF Edit). Figure 5 shows an excerpt of the classes generated for the implementation of the *Multipage* metamodel and an example of code usage for creating and manipulating a *Page* instance.

Furthermore, EMF Edit provides capabilities to build a graphical editor. It enables the visualization of model elements and their command-based editing. Figure 6 shows an overview of the architecture of a graphical editor build using the EMF Edit generated code. This will be needed in order to understand the solution exposed further in the paper since it extends the mechanism of EMF Edit. The generated code includes:

- *ItemProvider*(s): They are generated for each class of the metamodel. They are used to display model elements in a graphical editor via an Adapter design pattern (a delegation mechanism that makes it possible to "act as if" an object of type A was an object of type B).
- *ItemProviderAdapterFactory*: It is used to group all *ItemProvider* classes and provide a centralized mechanism to request them.
- *ContentProvider*(s) and *LabelProvider*(s): They are used to provide the display of an item. A *ContentProvider* retrieves the content of an item displayed by a graphical interface where the *LabelProvider* takes care of the visualization (image and text) of the item. The *ContentProvider*(s) and *LabelProvider*(s) can (and usually should) delegate to the same *AdapterFactory* and, therefore, to the same *ItemProvider*(s).
- *ComposedAdapterFactory*: It is useful in order to stick different adapter factories (for individual models).
- *EditingDomain*: It is an editing command structure, including a set of generic command implementation classes to build editors that fully support, cancel and redo actions.

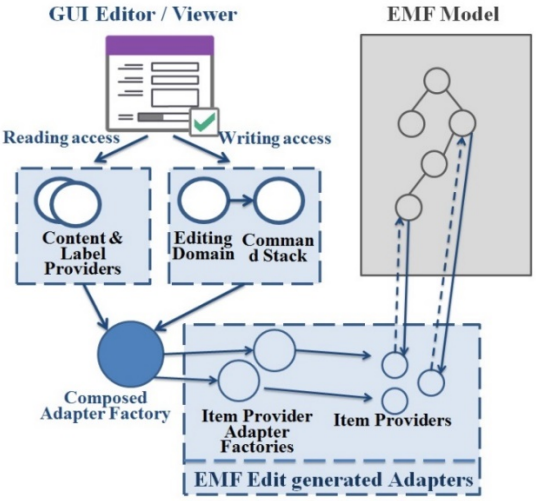


Figure 6. The architecture of a graphical editor built using EMF Edit.

Figure 7 shows a class diagram representing the EMF Edit generated code for the *Page* class from the *Multipage* metamodel.

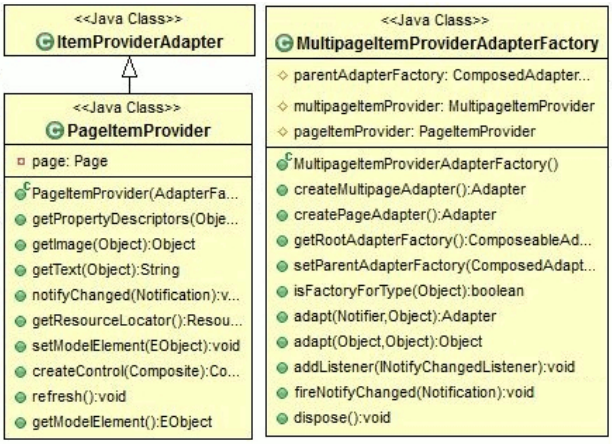


Figure 7. A class diagram representing EMF Edit generated classes for the metaclass Page.

8 shows an excerpt of the *Sheet* editor implementation. It is important to mention that the provision of the graphical content for each element of the model was centralized in a *creatCon-trol()* method attached to its adapter *ItemProvider*. The resulted editors match the screenshots shown in Figure 3.

The code generator facility

As mentioned earlier, the PoC aims to implement a code generator facility that allows generating a Java code overlay of the generated EMF Edit code. It must provide the necessary infrastructure to make it possible to quickly build (compose) new DSMLs based on the composition of their metamodels; using the composition rules defined in Subsection 4.2.

```
/**
 * This is an example of a Sheet model editor.
 */
public class SheetEditor extends EditorPart implements IResourceChangeListener, IResourceDeltaVisitor {
... ...
public void createPartControl(Composite parent) {
// This is the one adapter factory used for providing views of the model.
ComposedAdapterFactor adapterFactory = new ....
SheetItemProvider sItemItemprovider = new SheetItemProvider(adapterFactory);
adapterFactory.addAdapterFactory(sItemItemprovider);
adapterFactory.addAdapterFactory(...);
... ...
// Load the resource through the editing domain.
Sheet sheet = getEditingDomain().getResourceSet().getResource(resourceURI, true).getContents()....;
// Attach the model element to its item provider
sItemItemprovider.setModelElement(sheet);
... ...
// create the Sheet GUI component
adapterFactory.adapt(sheet, SheetItemProvider.class).createControl(parent);
}
... ...
}

/**
 * This is the item provider adapter for a Sheet object.
 * @generated
 */
public class SheetItemProvider extends PageItemProvider {

/** @generated not */
public Composite createControl(Composite parent) {
//Content & Label Provider initialisation
ITreeContentProvider contentProvider =  new SheetViewTreeContentProvider();
DelegatingStyledCellLabelProvider labelProvider = new DelegatingStyledCellLabelProvider(...);
Composite content = new Composite(parent, SWT.NONE);
... ...
// create the model element GUI component
TreeViewer treeViewer = new TreeViewer();
treeViewer.setContentProvider(contentProvider);
treeViewer.setInput((Sheet)getModelElement());

// Iterate over the sheet element to create columns and rows using Content and Label Provides
... ...
return content;
}
... ...
}
```

Figure 8. An excerpt from the code of the Sheet Editor.

Therefore, the adapters mechanism used by EMF Edit is extended with the definition of a Java interface called *IExtendedGraphicalItemProvider*. It specifies a contract of five methods sufficient to create a graphical component connected to an element of the model, to refresh it and to dispose it:

- *createControl()*: It centralizes the provision of the graphical content of the model element.
- *setModelElement()*: It attaches the model element to its provider.
- *getModelElement()*: It accesses the model element from the provider.
- *refresh()*: It refreshes the graphical content of the model element.
- *dispose()*: It disposes the graphical content of the model element.

The implemented code generator facility modifies each edit *ItemAdapter* class, already generated by EMF Edit, in order to make it extend the *IExtendedGraphicalItemProvider* interface. These methods must be implemented and used for the construction of a new graphical editor. So far nothing new compared to the classic use of EMF Edit adapters. Now, if a composition rule is applied between two metamodels, these extended adapters come into play with the use of the aforementioned methods.

Let us consider the following example to better illustrate these statements. Let MM_A and MM_B be two metamodels related to two DSMLs (α) and (β). Let M_A be a model conforming to MM_A and M_B be a model conforming to MM_B. Let $\otimes Rule_C$ be a composition rule MM_A and MM_B in order to create a new DSML (∂). Considering that $\otimes Rule_C$ implies that an element A of the model M_A has to be linked to an element B of the model M_B. A graphical editor built using the architecture shown in Figure 6 makes that the *ItemProviderAdapterFactor* calls the *ItemProvider* IP_A related to the element A to display it in the editor of (α). Likewise, the *ItemProviderAdapterFactor* calls the *ItemProvider* IP_B related to the element B to display it in the editor of (β). In a new architecture built using the extended code generator facility, the generated adapters IP_A and IP_B will be linked with a generated link which reflects the $\otimes Rule_C$ rule. Thus, IP_A will directly call IP_B for displaying element B in the editor of (∂). Figure 9 schematizes this new architecture. For example, a specialization rule will imply, at the generated code, inheritance between IP_A and IP_B. Whereas containment reference rule will imply encapsulation of methods (defined by the interface *IExtendedGraphicalItemProvider*) of IP_B by those of IP_A.

Demonstration of the generator

Let us return back to our case study to illustrate the extended code generator facility through a second example. Let us consider the composition rule $\otimes Spe_{page}$ outlined in Subsection 4.3. The rule was applied on the implemented Ecore metamodels *Sheet* and *Multipage*. Indeed, Ecore makes it possible to describe a link of specialization between two metaclasses of two different metamodels. Then, the EMF generator facility was used to generate the EMF Impl code and the EMF Edit code. Finally, the implemented code generator prototype was used to generate the extension layer with the *ItemProvider*(s) that extend the *IExtendedGraphicalItemProvider* interface. Figure 10 shows a class diagram representing the generated *ItemProvider*(s).

The generated code was used to re-implement the graphical editor of the composed DSML resulting of $\otimes Spe_{page}$. An extended *Multipage* editor has been obtained. It allows a multiple page to be composed of multiple sheet tabs. Figure 11 shows an excerpt from the code of the extended *Multipage* editor. It demonstrates how the *PageItemProvider* delegates the creation of the graphical component of an instance of *Sheet*, included under a *Multipage* instance, to a *SheetItemProvider*. It takes advantage of the polymorphism between the *PageItemProvider* and the *SheetItemProvider* to call the appropriate graphical interface creation method. In the same way, the code of the *MultipageItemProvider* demonstrates how the *refresh* and *dispose* methods can be called for each instance of *Page* contained in an instance of *Multipage*. It takes advantage of polymorphism to apply the appropriate method depending on whether the displayed instance is an instance of *Page* or an instance of *Sheet*.

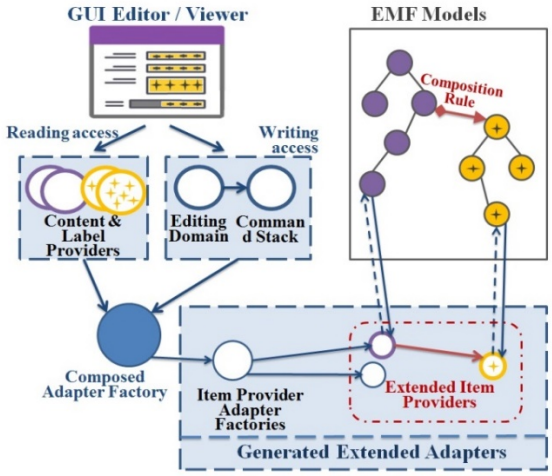


Figure 9. The architecture of a graphical editor built using the extended code generator facility.

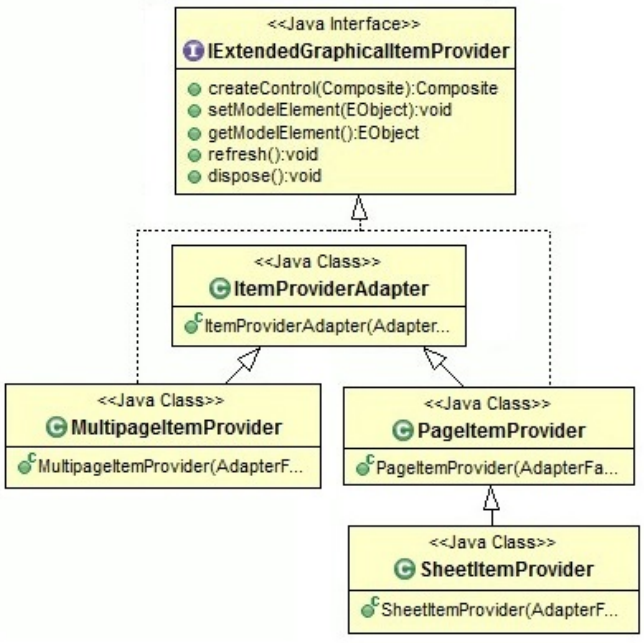


Figure 10. A class diagram representing the generated Item Providers using the extended code generator facility.

In this paper we have conducted an exploratory study whose goal is to explore means of composing DSMLs by composing their

```
/**
 * This is an example of an Extended Multipage model editor.
 */
public class ExtendedMultipageEditor extends EditorPart
                                     implements IResourceChangeListener, IResourceDeltaVisitor {
... ...
public void createPartControl(Composite parent) {
// This is the one adapter factory used for providing views of the model.
ComposedAdapterFactor adapterFactory = new ....
MultipageItemProvider mItemItemprovider = new MultipageItemProvider(adapterFactory);
adapterFactory.addAdapterFactory(mItemItemprovider);
adapterFactory.addAdapterFactory(...);
... ...
// Load the resource through the editing domain.
Multipage multipage = getEditingDomain().getResourceSet()...;
// Attach the model element to its item provider
mItemItemprovider.setModelElement(multipage);
... ...
// create the Multipage GUI component
tabFolder = (TabFolder) mItemItemprovider.createControl(parent);
  for (Page page : multipage.getChildren()) {
    String tabItemName = page.getName();
    TabItem tabItem = new TabItem(tabFolder, SWT.NONE);
    tabItem.setText(tabItemName);
    PageItemProvider pItemItemprovider  = (PageItemProvider) new SheetItemProviderAdapterFactory().
                                          getAdaptedProvider(page, adapterFactory);

    pItemItemprovider.setModelElement(page);

    // The Page Item Provider delegates the creation of the GUI omponent to a Sheet Item Provider
    // in order to create a Sheet GUI component
    Composite tabItemComposite = pItemItemprovider.createControl(tabFolder);
    tabItem.setControl(tabItemComposite);
  }
}
... ...
}
/**
 * This is the item provider adapter for a Multipage object.
 * @generated
 */
public class MultipageItemProvider extends ItemProviderAdapter
                                   implements IItemPropertySource, IItemLabelProvider,
                                              IStructuredItemContentProvider, ITreeItemContentProvider,
                                              IExtendedGraphicalItemProvider, EditingDomainItemProvider{
... ...
/**
 * @generated
 */
@Override
    public void refresh() {
    for (Page page : multipage.getChildren()) {
        // It delegates to the Sheet Item Provider if the page is an instance of Sheet
        adapterFactory.adapt(page, PageItemProvider.class).refresh(parent);
    }
/**
 * @generated
 */
@Override
    public void dispose() {
    for (Page page : multipage.getChildren()) {
        // It delegates to the Sheet Item Provider if the page is an instance of Sheet
        adapterFactory.adapt(page, PageItemProvider.class).dispose(parent);
    }
... ...
}
```

Figure 11. An excerpt from the code of the Extended Multipage Editor.

metamodels and studies the projection on their associated graphical editors. Three rules of composition were defined: Reference, Specification and Fusion. Each has been applied to a case study to illustrate its use. The impact of this composition on graphic editors has been studied. We have shown how the actual syntaxes of the originals DSMLs have been reused and composed to give shape to the concrete syntax of the composed DSML.

5.1. Development Time

Saving time is saving money. Development time is a major factor in software development. Indeed, with the multitude of technology and the vertiginous speed with which languages of programming develop. It is essential to minimize the development time of new software. It is one of the MDE's battle horses as it introduces the necessary abstraction to safeguard knowledge and automate the projection to technological spaces. In our exploratory study we have implemented a code generation prototype that allows taking advantage of our composition rules. It allows generating a layer of code that facilitates the composition of the concrete syntaxes of composed DSMLs. We have implemented it on our case study and we have shown an example in the previous section. After the method usage, we can measure about some preliminary results about the gain obtained by our approach in term of development time by measuring the percentage of generated of code. Because the percentage of code generated is directly correlated to the development time earned. Indeed, the less time spent writing the automatically generated code represents a time gained directly on the development time. In addition, we measure this value on both EMF Edit and our prototype. In this way we show what we also gain compared to the EMF Edit Framework.

Table 1 presents a comparison results summary. The first column lists the global number of line of code of each DSML. The second column compares the number of line of generated code by EMF Edit and by our generator facility. It worth to note that our generator is only used after a composition. Therefore, it has been used only for DSMLs (d), (e), (f) and (g). The third column shows a percentage comparing between the two tools. This last result is exploited in Fig. 12 to show by interpolation the potential gain in terms of development time. It is important to remember that our generator is an overlay of EMF Edit.

5.2. Code Ruse

One of our stated objectives in this study was the reuse of software components. We explored the reuse of existing DSMLs to extend or compose new ones. This reuse is reflected on the code of the obtained DSML. Thus we measured the percentage of reused code each time we extended our DSML in the application of the exploratory study to the case study.

Table 2 shows the percentage of code reuse each time we extended our DSMLs of the case study. It represents what we reused after applying each composition rule.

5.3. Learnability

In our case study, 100% of the graphic components of the original DSMLs were reused. Very few new graphical features have been introduced in composed DSMLs. This is a very important factor for the ease of learning of users. The learnability of software is often overlooked. However, it is the most influential aspect leading to the success of a software application.

In [78], authors noted that experience with similar software is a major dimension of learnability.

Table 1. Comparison between EMF Edit and Our Generator.

	lines of code	*generated Lines*		*% of generated code*	
		EMF Edit	Our Generator	EMF Edit	Our Generator
DSML(a)	1108	803		72%	
DSML(b)	2637	2369		90%	
DSML(c)	2735	1112		41%	
DSML(d)	3581	1532	3481	43%	97%
DSML(e)	3631	1532	3481	42%	96%
DSML(f)	4589	1916	4309	42%	94%
DSML(g)	4535	1901	4264	42%	94%

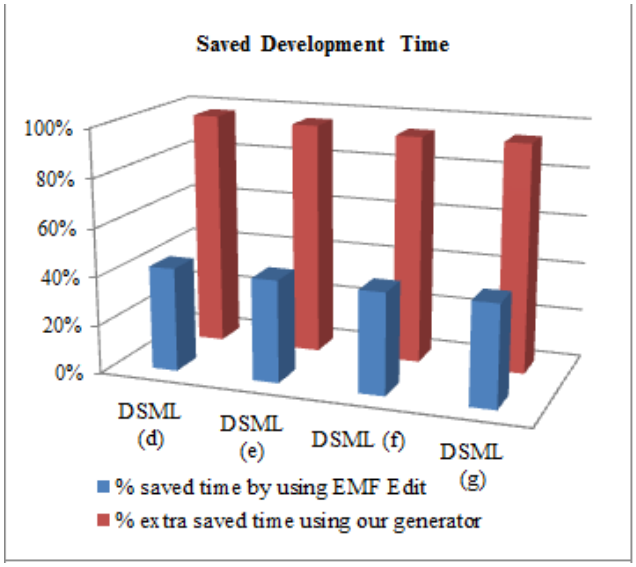


Figure 12. Gain in terms of development time.

Table 2. Percentage of reused after extending DSMLs.

	lines of code	*% of reused code*
DSML(d)	3581	97%
DSML(e)	3631	99%
DSML(f)	4589	97%
DSML(g)	4535	97%

6. Conclusions

This paper investigated the problem of extending DSMLs by composition their metamodels through an exploratory study. It exposes how DSMLs can be reused to rapidly create new ones with low cost. For this purpose three rules to compose DSMLs metamodels were specified: reference, specialization and fusion. A case study was used to illustrate the approach. In addition, the paper presented the implementation of a prototype of a code generator facility based on the aforesaid three composition rules. This prototype is then applied to the case study in order to validate our approach and measure its advantages. Compared to other works, our approach presents advantages, mainly by providing a higher level of reuse of DSMLs artefacts and by providing an automatic generation that facilitates the implementation of DSMLs tools and save development time. In addition, it keeps the graphics interfaces of the original DSMLs thus significantly improving the ease of learning of the new DSMLs.

The main contributions of the paper were: (i) the evaluation of the approach through an exploratory method; and (ii) the

implementation and the experimentation of the code generator facility prototype. Nevertheless, this work is only at its beginning. Indeed, it can be interesting to enlarge the set of composition rules, getting inspired by other principles and patterns coming from modeling languages and programming languages such as: encapsulation, substitution, adaptation and many others. Moreover, it can be interesting to take into account the composability properties of metamodels. Otherwise, the case study used in this study is very simple. It is a choice of writers to better illustrate the approach. However, it can exaggerate the results obtained from the fact of this simplicity.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We express our respected gratitude goes to Ibn Zohr University, LabSIV laboratory, Ibn Zohr Doctoral Study Center and Faculty of Sciences of Ibn Zohr University for funding this research..

References

[1] A. Abouzahra, A. Sabraoui, K. Afdel, "A Metamodel Composition Driven Approach to Design New Domain Specific Modeling Languages" in 1st European Conference on Electrical Engineering and Computer Science, Bern, Switzerland, 2017. https://doi.org/10.1109/EECS.2017.30

[2] T. Mens, "A State-of-the-Art Survey on Software Merging" IEEE Trans. Softw. Eng. 28(5), 449-462, 2002. https://doi.org/10.1109/TSE.2002.1000449

[3] S. Kent, "Model Driven Engineering" Lect. Notes. Comput. Sc., 2335, 286-298, 2002. https://doi.org/10.1007/3-540-47884-1_16

[4] D. C. Schmidt, "Guest Editor's Introduction: Model-Driven Engineering" Computer., 39(2), 25-31,2006.https://doi.org/10.1109/MC.2006.58

[5] R. Reddy, R. France, S. Ghosh, F. Fleurey, B. Baudry, "Providing Support for Model Composition in Metamodels" in 11th IEEE International Enterprise Distributed Object Computing Conference, Annapolis, MD, USA, 2007. https://doi.org/ 10.1109/EDOC.2007.55

[6] J. Estublier, G. Vega, A. D. Ionita, "Composing Domain-Specific Languages for Wide-Scope Software Engineering Applications" Lect. Notes. Comput. Sc., 3713, 69-83, 2005. https://doi.org/10.1007/11557432_6

[7] F. Fleurey, B. Baudry, R. France, S. Ghosh, "A Generic Approach for Automatic Model Composition" Lect. Notes. Comput. Sc., 5002, 7-15, 2008. https://doi.org/10.1007/978-3-540-69073-3_2

[8] J. Bézivin, R. F. Paige, U. Assmann, B. Rumpe, D. C. Schmidt, "Manifesto - Model Engineering for Complex Systems" Dagstuhl Seminar Proceedings, 08331, 2008.

[9] D. S. Kolovos, L. M. Rose, N. Matragkas, R. F. Paige, E. Guerra, J. S. Cuadrado, J. De Lara, I. Ráth, D. Varró, M. Tisi, J. Cabot, "A research roadmap towards achieving scalability in model driven engineering" in Workshop on Scalability in Model Driven Engineering (BigMDE '13), New York, USA, 2013.https://doi.org/10.1145/2487766.2487768

[10] B. Rumpe, "Towards model and language composition" in 1st Workshop on the Globalization of Domain Specific Languages (GlobalDSL '13), New York, USA, 2013.https:// doi.org/10.1145/2489812.2489814

[11] A. Horst, B., Rumpe, "Towards Compositional Domain Specific Languages" Ceur. Workshop. Procee., 1112, 7-16, 2013.

[12] U. Hohenstein and C. Elsner, "Model-driven development versus aspect-oriented programming a case study" in 9th International Conference on Software Paradigm Trends (ICSOFT-PT), Vienna, Austria, 2014. https://doi.org/ 10.5220/0004999901330144

[13] S. Dmitriev, Language oriented programming - the next programming paradigm, http://www.onboard.jetbrains.com/articles/04/10/lop/, accessed 21 November 2018.

[14] J.A. Pereira, K. Constantino, E. Figueiredo, "A Systematic Literature Review of Software Product Line Management Tools" Lect. Notes. Comput. Sc., 8919, 73-89, 2014. https://doi.org/10.1007/978-3-319-14130-5_6

[15] J. White, J. H. Hill, J. Gray, S. Tambe, A. S. Gokhale, D. C. Schmidt, "Improving Domain-Specific Language Reuse with Software Product Line Techniques" IEEE Software., 26(4), 47-53, 2009. https://doi.org/10.1109/MS.2009.95

[16] The Epsilon Homepage, https://www.eclipse.org/epsilon/, accessed 21 November 2018.

[17] The Epsilon Merging Language (EML) Homepage, http://www.eclipse.org/epsilon/doc/eml/, accessed 21 November 2018.

[18] D. S. Kolovos, R. F. Paige, F. A. C. Polack, "Merging models with the epsilon merging language (EML)" Lect. Notes. Comput. Sc., 4199, 215-229, 2006. https://doi.org/10.1007/11880240_16

[19] D. Kolovos, "Merging Models with the Epsilon Merging Language - A Decade Later", in 19th ACM/IEEE International Conference on Model Driven Engineering languages and Systems, Saint-Malo, France, 2016.

[20] M. D. Del Fabro, P. Valduriez, "Towards the efficient development of model transformations using model weaving and matching transformations" Softw. Syst. Model., 8(3), 305-324, 2009. https://doi.org/10.1007/s10270-008-0094-z

[21] The MOMENT web site, http://moment.dsic.upv.es//, accessed 21 November 2018.

[22] A. Boronat, "MOMENT: A Formal Framework for MOdel managemMENT" Ph.D Thesis, Universitat Politècnica de València, 2007.

[23] The Eclipse Modeling Framework (EMF) Homepage, http://www.eclipse.org/modeling/emf/, accessed 21 November 2018.

[24] A. Boronat, J. Á. Carsí, I. Ramos, "Automatic Support for Traceability in a Generic Model Management Framework" Lect. Notes. Comput. Sc., 3748, 316-330, 2005. https://doi.org/10.1007/11581741_23

[25] QVT, The MOF Query/View/Transformation specification page, http://www.omg.org/spec/QVT/, accessed 21 November 2018.

[26] A. Boronat, J. Á. Carsí, I. Ramos, P. Letelier, "Formal model merging applied to class diagram integration" Electron. Notes Theor. Comput. Sci., 166, 5-26, 2007. https://doi.org/10.1016/j.entcs.2006.06.013

[27] T. Degueule, B. Combemale, A. Blouin, O. Barais, J.M. Jézéquel, "Melange: a meta-language for modular and reusable development of DSLs" in 2015 ACM SIGPLAN International Conference on Software Language Engineering (SLE 2015), New York, USA, 2015. https://doi.org/10.1145/2814251.2814252.

[28] T. Degueule, B. Combemale, A. Blouin, O. Barais, J. M. Jézéquel, "Safemodel polymorphism for flexible modeling" Comput. Lang. Syst. Struct, 49(C),176-195, 2017. https://doi.org/10.1016/j.cl.2016.09.001

[29] S. Kelly, K. Lyytinen, M. Rossi, "Metaedit + a fully configurable multi-user and multi-tool case and came environment" Lect. Notes. Comput. Sc., 1080, 1-21, 1996. https://doi.org/10.1007/3-540-61292-0_1

[30] S. Kelly, J. P. Tolvanen, Domain-specific modeling: enabling full code generation , John Wiley & Sons, 2008.

[31] S. Erdweg, T. van der Storm, M. Völter, L. Tratt, R. Bosman, W. R. Cook, A. Gerritsen f, Angelo Hulshout g, Steven Kelly h, Alex Loh c, Gabriël Konat i, Pedro J. Molina j, Martin Palatnik, R. Pohjonen, E. Schindler, K. Schindler, R. Solmi, V. Vergu, E. Visser, K. van der Vlist, G. Wachsmuth, J. van der Woning, "Evaluating and comparing language workbenches: existing results and benchmarks for the future" Comput. Lang. Syst. Struct., 44(PA), 24-47, 2015. https://doi.org/10.1016/j.cl.2015.08.007

[32] H. Berg, B. Møller-Pedersen, "Type-Safe Symmetric Composition of Metamodels Using Templates" Lect. Notes. Comput. Sc., 7744, 160-178, 2012. https://doi.org/10.1007/978-3-642-36757-1_10

[33] H. Berg, B. Møller-Pedersen, "Metamodel and Model Composition by Integration of Operational Semantics" Comm. Com. Inf. Sc., 580, 172-189, 2015. https://doi.org/10.1007/978-3-319-27869-8_10

[34] Schmidt, M., Wenzel, S., Kehrer, T., Kelter, U., "History-based merging of models" in 2009 ICSE Workshop on Comparison and Versioning of Software Models (CVSM '09), Washington, USA, 2009. https://doi.org/10.1109/CVSM.2009.5071716

[35] H. K. Dam, A. Egyed, M. Winikoff, A. Reder, R. E. Lopez-Herrejon, "Consistent merging of model versions" J. Syst. Softw., 112(C), 137-155, 2016. https://doi.org/10.1016/j.jss.2015.06.044

[36] D. Zhang, S. Li, X. Liu, "An Approach for Model Composition and Verification" in 1th IEEE Computer Society International Joint Conference on INC, IMS and IDC, Seoul, South Korea, 2009. https://doi.org/10.1109/NCM.2009.271

[37] The Alloy Homepage. http://alloy.mit.edu, accessed 21 November 2018.

[38] D. Jackson, "a lightweight object modelling notation" ACM Trans. Softw. Eng. Methodol., 11(2), 256-290, 2002. https://doi.org/10.1145/505145.505149.

[39] The Eclipse Graphical Modeling Framework (GMF) Homepage, http://www.eclipse.org/modeling/gmf/, accessed 21 November 2018.

[40] The Eclipse Modeling Project (EMP) Homepage, http://www.eclipse.org/modeling/, accessed 21 November 2018.

[41] Á. Lédeczi, A. Bakay, M. Maroti, P. Volgyesi, G. Nordstrom, J. Sprinkle, G. Karsai, "Composing domain-specific design environments", Computer., 34(11), 44-51, 2001. https://doi.org/10.1109/2.963443

[42] Á. Lédeczi, G. Nordstrom, G. Karsai, P. Volgyesi, M. Maroti, "On metamodel composition" in 2001 IEEE International Conference on Control Applications (CCA'01), Mexico City, Mexico, 2001. https://doi.org/10.1109/CCA.2001.973959

[43] J. M. Jézéquel, "Model driven design and aspect weaving" Softw. Syst. Model., 7(2), 209-218, 2008. https://doi.org/10.1007/s10270-008-0080-5
[44] O. Barais, J. Klein, B. Baudry, A. Jackson, S. Clarke, "Composing multi-view aspect models" in 7th International Conference on Composition-Based Software Systems (ICCBSS 2008), Washington, USA, 2008. https://doi.org/10.1109/ICCBSS.2008.12 2008
[45] P. Sánchez, L. Fuentes, D. Stein, S. Hanenberg, R. Unland, "Aspect-oriented model weaving beyond model composition and model transformation" Lect. Notes. Comput. Sc., 5301, 766-781, 2008. https://doi.org/10.1007/978-3-540-87875-9_53
[46] A. Hovsepyan, S. Van Baelen, Y. Berbers, W. Joosen, "Specifying and Composing Concerns Expressed in Domain-Specific Modeling Languages", In M. Oriol, B. Meyer (Ed.), Objects, Components, Models and Patterns, Lect. Notes. Bus. Inf., 33, 116-135, 2009. https://doi.org/10.1007/978-3-642-02571-6_8
[47] M. P. Cardoso, T. Carvalho, J. G. F. Coutinho, W. Luk, R. Nobre, P. Diniz, Z. Petrov, "LARA: an aspect-oriented programming language for embedded systems" in 11th annual international conference on Aspect-oriented Software Development, New York, USA, 2012. https://doi.org/10.1145/2162049.2162071
[48] P. Pinto, T. Carvalho, J. Bispo, M. A. Ramalho, J. M. P. Cardoso, "Aspect composition for multiple target languages using LARA" Comput. Lang. Syst. Struct, 53, 1-26, 2018. https://doi.org/10.1016/j.cl.2017.12.003
[49] J. Whittle, P. Jayaraman, A. Elkhodary, A. Moreira, J. Araúj, "MATA: A unified approach for composing UML aspect models based on graph transformation" Lect. Notes. Comput. Sc., 5560, 191-237, 2009. https://doi.org/10.1007/978-3-642-03764-1_6
[50] J. Whittle, P. Jayaraman, "MATA: A Tool for Aspect-Oriented Modeling Based on Graph Transformation" Lect. Notes. Comput. Sc., 5002, 16-27, 2008. https://doi.org/10.1007/978-3-540-69073-3_3
[51] M. Schöttle, O. Alam, FP. Garcia, G. Mussbacher, J. Kienzle, "TouchRAM: a multitouch-enabled software design tool supporting concern-oriented reuse" in 13th International Conference on Modularity. New York, USA, 2014. https://doi.org/10.1145/2584469.2584475
[52] M. Schöttle, N. Thimmegowda, O. Alam, J. Kienzle, G. Mussbacher, "Feature modelling and traceability for concern-driven software development with TouchCORE" in 14th International Conference on Modularity, New York, USA, 2015. https:// doi.org/10.1145/2735386.2735922
[53] M. Voelter, "Language and IDE modularization, extension and composition with MPS" Lect. Notes. Comput. Sc., 7680, 383-430, 2013. https://doi.org/10.1007/978-3-642-35992-7_11
[54] M. Voelter, J. Warmer, B. Kolb, "Projecting a modular future" IEEE. Software., 32(5), 46-52, 2015. https://doi.org/10.1109/MS.2014.103
[55] M. Voelter, B. Kolb, T. Szabó, D. Ratiu, A. van Deursen, "Lessons learned from developing mbeddr: a case study in language engineering with MPS" Softw. Syst. Model., 17(66), 1-46, 2017. https://doi.org/10.1007/s10270-016-0575-4
[56] A. M. Şutîi, "MetaMod: a modeling formalism with modularity at its core" in 30th IEEE/ACM International Conference on Automated Software Engineering, Lincoln, USA, 2015. https://doi.org/10.1109/ASE.2015.29
[57] A. M. Şutîi, M. Van Den Brand, T. Verhoeff, "Exploration of modularity and reusability of domain-specific languages: an expression DSL in MetaMod" Comput. Lang. Syst. Struct., 51(C), 48-70, 2018. https://doi.org/10.1016/j.cl.2017.07.004
[58] D. H. Lorenz, B. Rosenan, "Cedalion: a language for language oriented programming" SIGPLAN. Not., 46(10), 733-752, 2011. https://doi.org/10.1145/2076021.2048123
[59] D. H. Lorenz, B. Rosenan, "Code reuse with language oriented programming" Lect. Notes. Comput. Sc., 6727, 167-182, 2011. https://doi.org/10.1007/978-3-642-21347-2_13
[60] D. H., Lorenz, B. Rosenan, "CEDALIONs Response to the 2016 Language Workbench Challenge", in LWC@SLE 2016 Language Workbench Challenge at the 2016 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications, Amsterdam, Netherlands, 2016.
[61] L. C. L. Kats, E. Visser, "The Spoofax language workbench: rules for declarative specification of languages and IDEs" SIGPLAN. Not., 45(10), 444-463, 2010. https://doi.org/10.1145/1932682.1869497
[62] M. Voelter, S. Benz, C. Dietrich, B. Engelmann, M. Helander, L. C. L. Kats Kats, E. Visser, G. Wachsmuth, DSL engineering: Designing, implementing and using domain-specific languages, dslbook.org, 2013.
[63] The Xtext Homepage, http://www.eclipse.org/Xtext/, accessed 21 November 2018.
[64] L. Bettini, Implementing domain-specific languages with Xtext and Xtend, Packt Publishing Ltd, 2016.
[65] H. Krahn, B. Rumpe, S. Völkel, "Monticore: Modular development of textual domain specific languages" Lect. Notes. Bus. Inf., 11, 297-315, 2008. https://doi.org/10.1007/978-3-540-69824-1_17
[66] H. Krahn, B. Rumpe, S. Völkel, "MontiCore: a framework for compositional development of domain specific languages" Int. J. Softw. Tools Technol. Transf., 12(5), 353-372, 2010. https://doi.org/10.1007/s10009-010-0142-1
[67] L. Pedro, V. Amaral, D. Buchs, "Foundations for a domain specific modeling language prototyping environment: A compositional approach" in 8th OOPSLA workshop on domain-specific modeling. In Companion to the 23rd ACM SIGPLAN conference on Object-oriented programming systems languages and applications (OOPSLA Companion '08). ACM, New York, USA, 2008. https://doi.org/10.1145/1449814.1449886
[68] L. Pedro, M. Risoldi, D. Buchs, B. Barroca, V. Amaral, "Composing Visual Syntax for Domain Specific Languages" Lect. Notes. Comput. Sc., 5611, 889-898, 2009. https://doi.org/10.1007/978-3-642-02577-8_97
[69] B. Meyers, "A Multi-Paradigm Modelling Approach for the Engineering of Modelling Languages" Ceur. Workshop. Procee., 1321, 2-9, 2015.
[70] B. Meyers, A. Cicchetti, E. Guerra, J. de Lara, "Composing textual modelling languages in practice" in 6th International Workshop on Multi-Paradigm Modeling, New York, USA, 2012. https://doi.org/10.1145/2508443.2508449
[71] J. De Lara, E. Guerra, "Deep metamodelling with metaDepth" Lect. Notes. Comput. Sc., 6141, 1-20, 2009. https://doi.org/10.1007/978-3-642-13953-6_1
[72] C. Herrmann, H. Krahn, B, Rumpe, M. Schindler, S. Völkel, "An Algebraic View on the Semantics of Model Composition" Lect. Notes. Comput. Sc., 4530, 99-113, 2007. https://doi.org/10.1007/978-3-540-72901-3_8
[73] S. Kelly, J. P. Tolvanen, Domain-Specific Modeling: Enabling Full Code Generation, Wiley, 2008.
[74] S. Völkel, "Kompositionale entwicklung domänenspezifischer sprachen," Ph.D Thesis, Technical University Carolo-Wilhelmina, 2011.
[75] The Xml Metadata Interchange (XMI) Specification page, http://www.omg.org/mof/ http://www.omg.org/spec/XMI/, accessed 21 November 2018.
[76] The MetaObject Facility (MOF) Specification page, http://www.omg.org/mof/, accessed 21 November 2018.
[77] The Eclipse Standard Widget Toolkit (SWT) Homepage, https://www.eclipse.org/swt/, accessed 21 November 2018.
[78] T. Grossman, G. W. Fitzmaurice, R. Attar, "A survey of software learnability: metrics, methodologies and guidelines" in 27th International Conference on Human Factors in Computing Systems, New York, USA, 2009. https://doi.org/10.1145/1518701.1518803

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 372-379 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Modeling an Energy Consumption System with Partial-Value Data Associations

Nong Ye*, Ting Yan Fok, Oswald Chong

School of Computing, Informatics and Decision Systems Engineering, Ira A. Fulton School of Engineering, Arizona State University, 85287-8809, USA

ARTICLE INFO

Article history:
Received: 13 September, 2018
Accepted: 20 November, 2018
Online: 05 December, 2018

Keywords:
Partial-Value Association,
Data Mining,
Energy Consumption,
Structural System Model

ABSTRACT

Many existing system modeling techniques based on statistical modeling, data mining and machine learning have a shortcoming of building variable relations for the full ranges of variable values using one model, although certain variable relations may hold for only some but not all variable values. This shortcoming is overcome by the Partial-Value Association Discovery (PVAD) algorithm that is a new multivariate analysis algorithm to learn both full-value and partial-value relations of system variables from system data. Our research used the PVAD algorithm to model variable relations of energy consumption from data by learning full-and partial-value variable relations of energy consumption. The PVAD algorithm was applied to data of energy consumption obtained from a building at Arizona State University (ASU). Full- and partial-value variable associations of building energy consumption from the PVAD algorithm are compared with variable relations from a decision tree algorithm applied to the same data to show advantages of the PVAD algorithm in modeling the energy consumption system.

1. Introduction

Our research is an extension of work originally presented in the 2018 IEEE ICCAR Conference [1]. Many complex systems, such as energy consumption systems and transportation systems, involve both engineered and non-engineered system factors. For example, the energy consumption system of a building involves both engineered system factors, (e.g., AC equipment, pump for water use, lighting system, computers, and network equipment) and non-engineered system factors, (e.g., social/behavioral factors, such as occupants' activities, and environmental/natural factors such as outside climate), which are intertwined to drive the energy consumption and demand of the building [2-4]. For another example, the transportation system involves both engineered system factors, (e.g., the transportation infrastructure including highways, streets and roads, and traffic control mechanisms, such as traffic lights) and non-engineered system factors, (e.g., social/behavioral factors such as traffic flows, drivers, pedestrians, and car accidents, as well as natural/environmental factors, such as weather conditions).

Although models of engineered systems may be available, models of mixed-factor systems are usually not available due to unknown interconnectivities and interdependencies of many engineered and non-engineered system factors. A complete, accurate system model, which clearly defines relations of system variables including interconnectivities and interdependencies of engineering and non-engineered system factors, is highly desirable for many applications. For example, variable relations of energy consumption are required to enable the accurate estimation of energy consumption/demand and the close alignment of energy production with energy demand to achieve energy production and use .

Utility/energy companies currently rely heavily on the past data of electricity loads in base, average and peak to project energy production/supply. This statistical investigative activity is done without adequate and accurate models of energy consumption systems [3]. Power plants often generate enough power to satisfy base loads and meet the difference between peak and base loads, sudden demand surge or any gap of energy supply and demand through their excess production capacities or by procuring from other energy sources [5, 6]. Historical data lack critical real-time features (e.g., the lag effect of historical data, and lack of finer levels and finer divisions in time and space) for the accurate projection and estimation of energy demand and consumption. Without adequate and accurate models of energy consumption systems, it is extremely difficult to obtain an accurate projection and estimation of energy demand and consumption. As a result,

* Nong Ye, Arizona State University, Email: nongye@asu.edu

energy has to be produced in excess in order to meet potential rise in demand. Energy production in excess is a significant cause of waste and inefficiency. Even with current technologies to obtain dynamic data of energy consumption systems in real time, the lack of adequate and accurate energy system models renders real-time dynamic system data useless for closely aligning energy production with energy demand to achieve energy production efficiency and energy use reduction. The ultimate energy efficiency through smart energy production and use will enable a shift from the existing code-, standard- and experience-based forecasting approach to a more dynamic, real-time and smart technology environment based on real-time data, models and analytics for the real-time, accurate estimation of energy consumption and smart technologies to align energy production with energy demand closely for energy use reduction and energy production efficiency.

Many statistical modeling, data mining and machine learning techniques for system modeling, including decision trees, regression analysis, artificial neural network, and Bayesian networks, have been used to analyze and model energy consumption and efficiency of equipment, homes and buildings [7-16]. System modeling techniques based on many existing statistical analysis, machine learning and data mining have a shortcoming of building variable relations for the full ranges of variable values using one model, although certain variable relations may hold for only some but not all variable values. This shortcoming is overcome by the PVAD algorithm that is a new multivariate analysis algorithm to learn both full-value and partial-value relations of system variables from system data. Our research used the PVAD algorithm to model variable relations of energy consumption from data by learning full-and partial-value variable relations of energy consumption. The PVAD algorithm was applied to building energy consumption data at ASU.

2. Shortcomings of existing techniques of system modeling from data

Existing methods of learning system models from data include statistical analysis [17-24] and data mining techniques [23-32]. With system modeling from data, classification and prediction can be performed to explain or find relations among system variables. Depending on the nature of data, there are several methods to analyze data using statistical techniques such as parametric, non-parametric and logistic regression. For example, when modeling categorical dependent variables, logistic regression can be applied [17, 21, 22]. In addition to decision and regression trees [23, 24], random forest and support vector machine are also considered [25-28, 29, 31]. However, the above methods assume that the role of a variable in a variable relation is known (i.e., which variable is an independent or dependent variable) and a variable plays only one role of being either an independent variable or a dependent variable in one layer of variable relations. Once a variable is considered as an independent variable, it can no longer be utilized as a dependent variable which is a main disadvantage especially when the role of a variable is not known or when multiple layers of variable relations are required where a variable can play different roles of being an independent or dependent variable in different variable relations at different layers.

Bayesian networks [23, 24, 35-37], structural equation models [33, 34] and reverse engineering methods [38-47] are examples of a few options left that can provide system modeling without prior knowledge of variables. However, those techniques discover only variable relations for full ranges of all variable values instead of relations for specific values only. This can be seen from the Fisher's Iris data set [48] in which the classification of the target variable (Plant Type) using independent variables works for only the values of Iris Versicolor and Iris Virginica) for the target variable but not for another target value of Iris Sentosa. For such data where variable relations hold for partial ranges of variable values only or different variable relations hold for different ranges of variable values, the model of the same variable relations for all variable values do not fit all data values well, that is, the model explains or represents the whole data set poorly.

The PVAD algorithm was developed as a new system modeling technique [49-51] to overcome the above shortcomings. Variable value associations can be used to construct associative networks as multi-layer structural system models. The application of the PVAD based system modeling technique is part and parcel of our research of energy consumption in systems.

3. The energy consumption data and the PVAD application

The PVAD algorithm is presented in detail in [49-51]. This section shows the PVAD application to data of energy consumption collected from an ASU building in January 2013 for modeling energy consumption. There was a data sample every 15 minutes. The data set has 2976 data records or instances. Each data record contains four numeric values for the consumption of electricity (E), cooling (C), heating (H), and air temperature (A), respectively, as well as TimeStamp (T). T is important because changes of T are associated with changes in presence and activities of occupants and changes of E, C and H.

To apply the PVAD algorithm, in Step 1 the numeric variables of A, H, C, and E, were transformed into categorical variables as shown in Fi To apply the PVAD algorithm, in Step 1 the numeric variables of A, H, C, and E, were transformed into categorical variables as shown in Figures (1)-(4). More details of Step 1 are in [1].

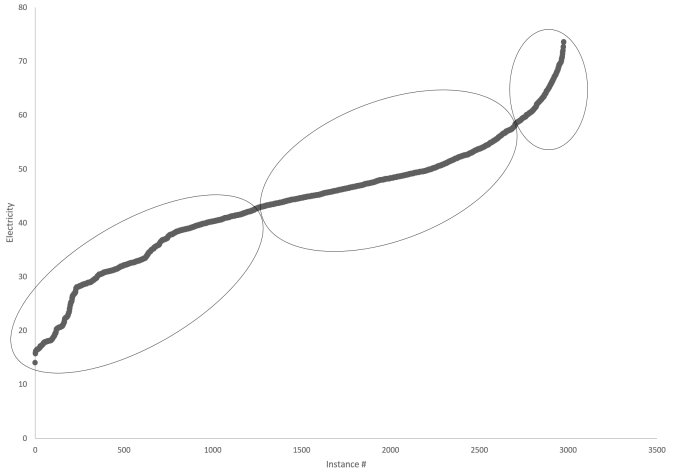


Figure 1. An example of plotting E values to determine data clusters and categorical values.

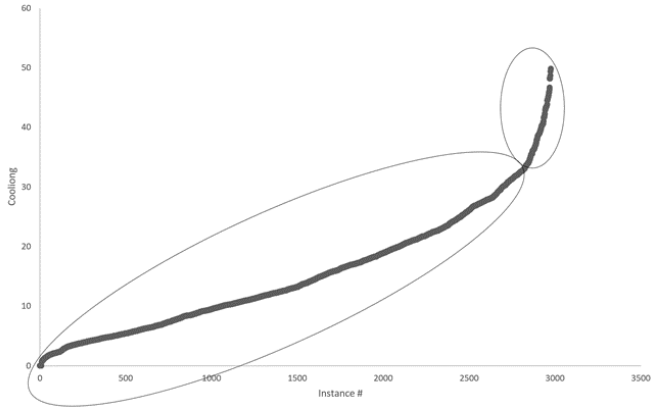
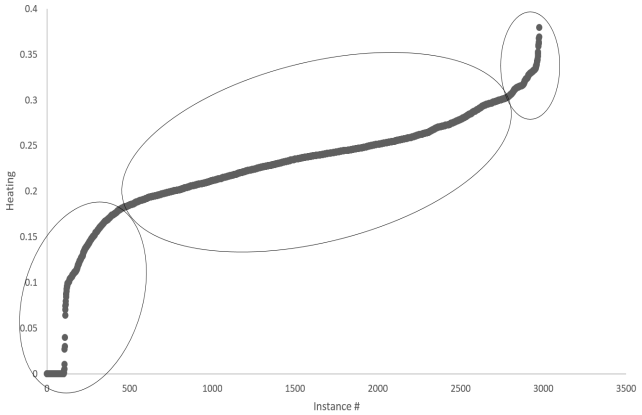


Figure 2. An example of plotting C values to determine data clusters and categorical values.



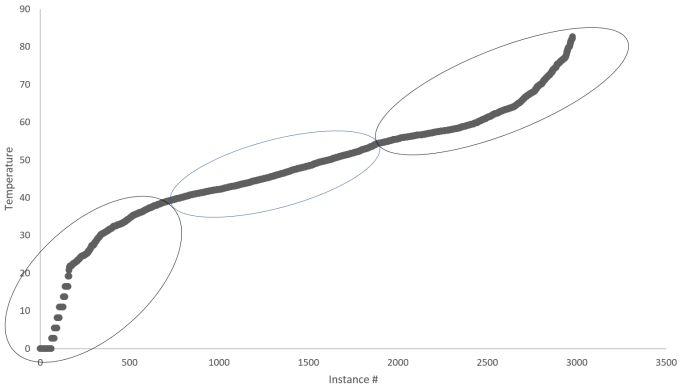


Figure 4. An example of plotting A values to determine data clusters and categorical values.

Step 2.1 generated candidate 1-to-1 associations of partial variable values, $x = a \rightarrow y = b$, where $x = a$ is the conditional variable value (CV) and $y = b$ is the associative variable value (AV), and computed the co-occurrence ratio (cr) of each candidate association as follows:

$$cr(x = a \rightarrow y = b) = \frac{N_{x=a,y=b}}{N_{x=a}} \qquad (1)$$

If cr is greater than or equal to the parameter α, we had an established association. For example, Table (1) shows 1-to-1 associations having CV: C = High together with their respective cr values and $\alpha = 0.8$.

Table 1: 1-to1 associations of ASU energy consumption data with CV: C=High

#	Association	*cr*	Co-Occurrence Frequency ($N_{x=a,y=b}$)	Type of Association
1	C=High → T=12:15 PM to 5:30 PM	0.48	62	Candidate
2	C=High → T=5:45 PM to 11 PM	0.43	55	Candidate
3	C=High → T=8:15 AM to 12 PM	0.09	11	Candidate
4	C=High → E=High	0.20	25	Candidate
5	**C=High → E=Medium**	**0.80**	**103**	**Established**
6	**C=High → H=Low**	**0.73**	**94**	**Established**
7	C=High → H=Medium	0.27	34	Candidate
8	**C=High → A=High**	**0.96**	**123**	**Established**
9	C=High → A=Medium	0.04	5	Candidate

In addition to parameter α, two other parameters, β and γ, are also needed. β is used to remove associations whose number of supporting instances (the instances containing variable values in the numerator of equation 1) is smaller than β. γ is used to remove an association with a common CV or AV that appears in more than γ of the data set. In this example, β is set to be 50 while α and γ are set to 0.8 and 0.95, respectively.

Step 2.2 uses two methods, YFM1 and YFM2, to examine and esyablish p-to-q associations, $X = A \rightarrow Y = B$, where X and Y represent multiple variables. For example, using #5, 6 and 8 in Table (1), we applied YFM1 which considers all combinations of AVs covered in those associations so as to find 1-to-q associations, where q >1. To find 1-to-2 established associations, we first computed $N_{CV} = 103 \div 0.8046875 = 128$. Then we considered all possible combinations of two-variable AVs from the established 1-to-1 associations:

1. C=High->E=Medium, H=Low (from #5 and #6)
2. C=High->E=Medium, A=High (from #5 and #8)
3. C=High->H=Low, A=High (from #6 and #8).

For each 1-to-2 candidate associations above, $N_{\text{CommonSubset}}$, the number of instances in the common subset of supporting instance, was computed to calculate cr for the 1-to-2 association. The results are given in Table (2). In this case, C=High->H=Low, A=High is the only established association.

Table 2: Calculation for 1-to-2 associations with CV: C=High

#	Association	$N_{\text{CommonSubset}}$	*cr* (= $N_{\text{CommonSubset}}$ / 128)
1	C=High → E=Medium, H=Low	87	0.6796875
2	C=High → E=Medium, A=High	103	0.125
3	**C=High → H=Low, A=High**	**94**	**0.8046875**

YFM2 is used to find 2-to-1 associations. YFM2 considers all candidate associations (cr value in (0, 1]) not just established associations ($cr \geq \alpha$). Table (3) is used to illustrate YFM2 in the following.

1) Determine $m_i = \lceil n_i \times \alpha \rceil$. If we pick C=High->T=12:15 PM to 5:30 PM to start with, $n_1 = 62$. Then $m_1 = \lceil n_1 \times \alpha \rceil = 50$ and the 2-to-1 association that we would like to generate will have C=High, T=12:15 PM to 5:30 PM as CV. Note that if $m_i < \beta$, the whole group is dropped as the number of instances covered by the new CV is just the occurrence frequency which should be $\geq \beta$.

2i) Iterate through all other associations from Table (2). Skip

immediately to the next line if the AV of that association is the same as one picked in the previous step. For example, #2 has AV: T=5:45 PM to 11 PM that represents Timestamp. While the AV of the association also represents timestamp (T=12:15 PM to 5:30 PM), we skip to #3 without looking at the intersection of the instances.

2ii) Generate 2-to-1 association if $n_{\text{intersection}} \geq 50$. Table (3) lists the $n_{\text{intersection}}$ and the corresponding *cr* value.

Table 3: Calculation for 2-to1 associations of ASU energy consumption data with CV "C=High, T=12:15 PM to 5:30 PM"

Combination (Type of Association)	Association	$n_{\text{intersection}}$	*cr* (= $n_{\text{intersection}}/62$)
1 & 4 (Candidate)	C=High, T=12:15 PM to 5:30 PM → E=High	3	0.04
1 & 5 (Established)	**C=High, T=12:15 PM to 5:30 PM → E=Medium**	**59**	**0.9516**
1 & 6 (Candidate)	C=High, T=12:15 PM to 5:30 PM → H=Low	59	0.9516
1 & 7 (Candidate)	C=High, T=12:15 PM to 5:30 PM → H=Medium	3	0.04
1 & 8 (Established)	**C=High, T=12:15 PM to 5:30 PM → A=High**	**62**	**1**
1 & 9 (Candidate)	C=High, T=12:15 PM to 5:30 PM → A=Medium	0	0

Following the same procedure, other *p*-to-*q* associations were generated by YFM1 and YFM2. Step 3 generalized and consolidated variable associations of partial values into associations of full value ranges if there are partial-value associations covering the full value range of the same variable.

The PVAD algorithm was used to analyze the energy consumption data using various values of α = 1, 0.9, and 0.8, β = 50, 30, and 10, and γ = 95%. The results for γ = 95%, β = 50, and α = 0.8 are most meaningful and presented in the next section.

4. Results of the PVAD Algorithm

Tables (4)-(5) list the most specific association(s) in each group of the associations with the same AV. Table (6) lists the most generic association(s) in each group of the associations with the same AV. Variable relations for energy consumption revealed by each association in Tables (4)-(6). In Tables (4)-(6), there are groups that give similar associations. For example, the associations in Group 1 and Group 2 in Table (6) are similar. For the groups with similar associations, we marked only one group using the symbol ^ in the column of group #. Most of the associations in Tables (4)-(6) involve C=Low for cooling being low in CV or AV, because most of instances in the data set (2848 out of totally 2976 instances) contain C=Low due to the month of January when the data was collected. Since C=Low is so common in the data set, C=Low can be dropped from the associations when interpreting associations.

Table 4: Specific associations in each group of associations with the same AV: Set 1

Group #	The most specific association(s) in group
1	A=Medium, [T=12:15 PM to 11 PM, E=High]/ [T=6:15 AM to 11 PM, E=Medium]/[T=11:15 PM to 6 AM, E=Low] → H=Medium, C=Low
2^	A=Medium, C=Low, [T=11:15 PM to 6 AM, E=Low]/[T=6:15 AM to 12 PM, E=Medium]/[T=12:15 PM to 11 PM, E=High/Medium] → H=Medium A=High, C=Low, E=Medium, T=8:15 AM to 12 PM → H=Medium
3^	A=High, E=Medium, C=High, T=12:15 PM to 5:30 PM → H=Low
4	C=High, E=Medium, T=12:15 PM to 5:30 PM → A=High | H=Low
5^	H=High, E=Medium, T=6:15 AM to 8 AM → A=Low, C=Low
6	H=Medium, E=Low, T=12:15 PM to 5:30 PM → A=High, C=Low
7^	[E=Medium, C=*, H=Low]/[E=Low, C=Low, H=Medium], T=12:15 PM to 5:30 PM → A=High C=High, T=5:45 PM - 11 PM → A=High
8	H=Medium, E=High, T=12:15 PM to 5:30 PM → A=Medium, C=Low
9^	H=Medium, C=Low, E=High, T=12:15 PM to 5:30 PM, → A=Medium
10	H=High, C=Low, E=Medium, T=6:15 AM to 8 AM, → A=Low
11^	[A=Low, H=High]/[A=High, H=Low]/[A=Low/Medium, H=Medium], C=Low, T=11:15 PM to 6 AM → E=Low
12	[A=Low, H=High]/[A=High, H=Low]/[A=Medium/Low, H=Medium], T=11:15 PM to 6 AM → C=Low, E=Low

Table 5: Specific associations in each group of associations with the same AV: Set 2

13	A=Medium, H=High, T=5:45 PM to 11 PM → C=Low, E=Medium T=8:15 AM to 12 PM, H=Medium, A=High/Medium → E=Medium, C=Low
14^	A=High, H=Low, C=High, T=12:15 PM to 5:30 PM → E=Medium A=Medium, H=High, C=Low, T=5:45 PM to 11 PM → E=Medium A=Medium/High, H=Medium, C=Low, T=8:15 AM to 12 PM → E=Medium
15	H=Low, C=High, T=12:15 PM to 5:30 PM → A=High | E=Medium
16^	A=High, C=High, T=12:15 PM to 5:30 PM, → H=Low | E=Medium A=Medium/High, C=Low, T=8:15 AM to 12 PM → H=Medium | E=Medium

The associative network of the energy consumption system model shown in Figure (5) was constructed using the associations in the groups marked with ^ in Table (6). Figure (5) shows the factors associated with the high, medium and low air temperatures (from the associations with A as the AV), the factors associated with the Medium and Low heating consumption (from the associations with H as the AV), and the factors associated with the medium and low electricity consumption (from the associations with E as the AV).

Figure (5) shows that E, C, H and A are related differently in different time periods. For example, in the afternoon, T = 12:15 PM to 5:30 PM, the medium heating consumption (H = Medium) along with the high electricity consumption (E = High) is associated with the medium air temperature (A = Medium), whereas in the early morning, T = 6:15 AM to 8 AM, the high heating consumption (H = High) is associated with the low air temperature (A = Low). Similarly, the most specific associations

in Tables (4)-(5), even the most generic associations in Table (6) and in Figure (5) show that associations of T, E, C, H and A differ in different value ranges of these variables. This illustrates that the PVAD algorithm can discover full/partial-value variable relations that exist in many real-world systems.

Table 6: Generic associations in each group of associations with the same AV.

Group #	The most generic association(s) in each group
1^	A=Medium/E=High → H=Medium
2	E=High/A=Medium, C=Low → H=Medium
3	E=Medium, C=High, A=High → H=Low
4	E=Medium, C=High → H=Low | A=High
5^	H=High, T=6:15 AM to 8 AM → A=Low.
6^	E=Low, T=12:15 PM to 5:30 PM → A=High
7^	H=Low → A=High T=5:45 PM - 11 PM, C=High → A=High C=Low, E=Low, T=12:15 PM to 5:30 PM → A=High
8^	H=Medium, E=High, T=12:15 PM to 5:30 PM → A=Medium
9	H=Medium, C=Low, E=High, T=12:15 PM `to 5:30 PM → A=Medium
10	H=High, C=Low, T=6:15 AM to 8 AM → A=Low
11	C=Low, T=11:15 PM to 6 AM → E=Low
12^	T=11:15 PM to 6 AM → E=Low
13^	T=8:15 AM to 12 PM → E=Medium H=High, T=5:45 PM to 11 PM → E=Medium
14	H=High, C=Low, T=5:45 PM to 11 PM → E=Medium C=High → E=Medium
15	C=High → A=High | E=Medium
16^	A=High, C=High, T=12:15 PM to 5:30 PM → H=Low | E=Medium A=Medium/High, C=Low, T=8:15 AM to 12 PM → H=Medium | E=Medium

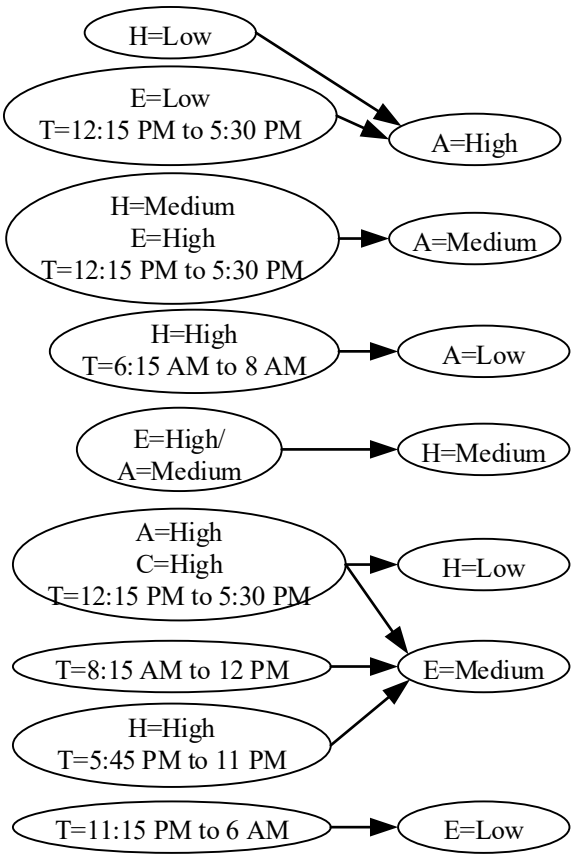


Figure 5. The most generic associations in the groups marked by ^ in Table (6) represented in an associative network.

5. Comparison of the PVAD algorithm with some data mining techniques

We considered two of the existing data mining techniques to compare with the PVAD algorithm: association rule and decision tree.

5.1. Comparison with the association rule technique

The association rule technique first uses the Aprori algorithm to determine frequent item sets that satisfy the minimum support [23-24]. Then each frequent item set is broken up into all possible combinations of association rules which are evaluated to see if any of them satisfy the minimum support and confidence. For a large dataset, frequent item sets and candidate association rules from frequent item sets can be enormous, requiring huge amounts of computer memory space and computation time. When the association rule technique was applied to the energy consumption data, there were too many frequent item sets and consequently association rules to be listed in this paper. While the performance of the association rule technique was hindered by the data size, the search space of associations in the PVAD algorithm is narrowed down by YMF1 and YFM2, along with parameters α, β and γ.

5.2 Comparison with the decision tree technique

Decision tree is a data mining technique to learn decision rules that express relations of the dependent variable y with independent variables $\boldsymbol{x}$ in a directed and acyclic graph [23-24]. The software, Weka, was used to construct decision trees of the energy consumption system data, To construct a decision tree in Weka, there are different algorithms such as ID3 [52] and J48 [53]. The later one is an extended version of ID3 with additional features like dealing with missing values and continuous attribute value ranges. It also addresses the over-fitting problem that decision trees are prone to by pruning. The pruning process requires the computation of the expected error rate. If the error rate of a subtree is greater than that of a leaf node, a subtree is pruned and replaced by the leaf node.

In our research, ID3 was used for the comparison with the PVAD algorithm because ID3 produces comparable results with associations produced by the PVAD algorithm. Leaf nodes produced by ID3 are pure in that the class labels of instances are the same in each leaf node. The purity of leaf node corresponds to AV in associations from the PVAD algorithm having the same variable value. The PVAD algorithm produces all associations up to N-to-1 associations, where N+1 is the number of variables. In other words, the PVAD algorithm can generate the longest CVs and find the AV that they are associated with. The combination of CVs corresponds to the path from the root of a decision tree down to a leaf node.

Because the decision tree technique requires the identification of one dependent variable (the target variable) and independent variables (attribute variables) for each decision tree, five decision trees need to be constructed for each of the five variables as the dependent variable. Tables (7)-(10) list decision rules produced by one of the five ID3 trees.

Although the decision rules from the decision trees appear to have the same form as associations from the PVAD algorithm, a

decision rule has a different meaning from an association from the PVAD algorithm. A decision rule derived from the root of a decision tree to a leaf node of the decision tree represents a frequent item set with instances in the leaf node having the values of the target variable and the attribute variables in the decision rule. This is why we see a path in a decision tree is also present in another tree even though different decision trees have different target variables. For example, the variable values in E=Medium, A=High, H=Medium, C=Low, T=12:15 PM to 5:30 PM, are found in all four decision trees. Note that the energy consumption data set has only five variables. Redundant paths of different decision trees can be found more often for larger data sets with more variables. This means the waste of computation time and space and the difficulty of sorting out results from a number of decision trees. Hence, a decision rule corresponds to a frequent item set in the association rule technique, whereas an association from the PVAD algorithm corresponds to an association rule in the association rule technique. This is why there are decision rules in Tables (7) – (10) that are not found in associations of the PVAD algorithm because frequent item sets for those decision rules were eliminated in the process of forming associations. Hence, the PVAD algorithm has the advantage to the decision tree technique because the PVAD algorithm discovers associations rather than frequent item sets.

Table 7: Decision rules from the ID3 Tree with Air Temperature as the target variable same as PVAD association rules

#	**Decision rules that appear same as PVAD association rules**
1	H=Medium, E=High, T=12:15 PM to 5:30 PM → A=Medium
2	H=Medium, E=Low, T=12:15 PM to 5:30 PM → A=High
3	H=Low, T=12:15 PM to 5:30 PM → A=High
4	H=High, E=Medium, T=6:15 AM to 8 AM → A=Low

Table 8: Decision rules from the ID3 Tree with Air Temperature = Low as the target variable

5	H=Medium, E=High, T=11:15 PM to 6 AM → A=Low
6	H=High, E=Low, T=11:15 PM to 6 AM → A=Low
7	H=High, E=Medium, T=11:15 PM to 6 AM →A=Low
8	H=High, E=Medium, T=12:15 PM to 5:30 PM → A=Low
9	H=High, E=Low, T=6:15 AM to 8 AM → A=Low
10	H=Medium, E=Low, T=6:15 AM to 8 AM → A=Low
11	H=High, E=Low, T=8:15 AM to 12 PM → A=Low
12	H=High, E=Medium, T=8:15 AM to 12 PM → A=Low
13	H=Low, C=Low, E=Medium, T=5:45 PM to 11 PM → A=Low

Table 9: Decision rules from the ID3 Tree with Air Temperature = Medium as the target variable

14	H=Low, T=6:15 AM to 8 AM → A=Medium
15	H=Medium, E=Low, T=11:15 PM to 6 AM → A=Medium
16	H=Low, E=Medium, T=11:15 PM to 6 AM → A=Medium
17	H=Medium, E=Medium, T=11:15 PM to 6 AM → A=Medium
18	H=High, E=Low, T=12:15 PM to 5:30 PM → A=Medium
19	H=High, E=Low, T=5:45 PM to 11 PM → A=Medium
20	H=High, E=Medium, T=5:45 PM to 11 PM → A=Medium
21	H=Medium, E=Medium, T=6:15 AM to 8 AM → A=Medium
22	H=Medium, E=High, T=8:15 AM to 12 PM → =Medium
23	H=Medium, E=Low, T=8:15 AM to 12 PM → A=Medium
24	H=Medium, C=Low, E=High, T=5:45 PM to 11 PM → A=Medium
25	H=Medium, C=Low, E=Medium, T=5:45 PM to 11 PM → A=Medium
26	H=Medium, C=Low, E=Medium, T=8:15 AM to 12 PM → A=Medium

Table 10: Decision rules from the ID3 Tree with Air Temperature = High as the target variable

27	H=Low, E=Low, T=11:15 PM to 6 AM → A=High
28	H=Low, E=High, T=5:45 PM to 11 PM → A=High
29	H=Low, E=Low, T=5:45 PM to 11 PM → A=High
30	H=Low, E=Low, T=8:15 AM to 12 PM → A=High
31	H=Medium, C=High, E=High, T=5:45 PM to 11 PM → A=High
32	H=Medium, C=Low, E=Low, T=5:45 PM to 11 PM → A=High
33	H=Low, C=High, E=Medium, T=5:45 PM to 11 PM → A=High
34	H=Medium, C=High, E=Medium, T=5:45 PM to 11 PM → A=High
35	C=High, H=Low, E=Medium, T=8:15 AM to 12 PM → A=High
36	H=Medium, C=High, E=Medium, T=8:15 AM to 12 PM → A=High
37	H=Low, C=Low, E=Medium, T=8:15 AM to 12 PM → A=High
38	H=Medium, C=High, E=Medium, T=12:15 PM to 5:30 PM → A=High
39	H=Medium, C=Low, E=Medium, T=12:15 PM to 5:30 PM → A=High

There is another difference between the decision tree technique and the PVAD algorithm. Each step of constructing a decision tree performs the splitting of a data subset for data homogeneity based on the comparison of splits using only one variable and its values rather than combinations of multiple variables due to the large number of combinations and the enormous computation costs. Hence, the resulting decision tree contains decision rules with the consideration of only one variable at a time and may miss decision rules that can be generated if multiple variables and their values are considered and compared at a time. However, the PVAD algorithm examines one to multiple variables at a time and does not miss any associations that exist. The PVAD algorithm thus has the advantage to the decision tree technique by not missing any established associations and using YFM1 and YFM2 to cut down the computation costs.

Moreover, the decision tree algorithm requires the identification of the dependent variable (the target variable) and the independent variables (the attribute variables) although there may no priori knowledge for the identification of which variable is a dependent or independent variable. This is why five decision trees, with one decision tree taking each of the five variables as the target variable, had to be constructed for the energy consumption data. The PVAD algorithm does not require the distinction of dependent and independent variables but discovers variable value relations and the role of each variable in each variable value relation.

Furthermore, the PVAD algorithm can generate *p*-to-*q* associations with $q > 1$ that the decision tree technique cannot generate because a decision tree is constructed for only one target variable and produces only *p*-to-1 decision rules. Given the differences of the PVAD algorithm and the decision tree technique, the results of the PVAD algorithm are not comparable to the results of the decision tree technique. As discussed in Section 2, the PVAD algorithm overcomes shortcomings of existing statistical analysis and data mining techniques and produce partial/full-value associations that cannot be produced from other existing techniques.

5. Conclusion

Our research used the PVAD algorithm to learn and build the system model of energy consumption from data, especially learn relations of variables for both full and partial value ranges. The resulting partial-value associations of variables in the energy consumption system model reveal variable relations for partial value ranges that require not one but different models of variable relations over full value ranges of the variables. This finding shows that the PVAD algorithm has the advantage and capability of discovering variable relations for building a multi-layer, structural system model. Hence, the PVAD based system modeling technique can be useful in many fields to learn system models from data. The advantages of the PVAD algorithm to existing data mining, machine learning and statistical analysis techniques were also demonstrated by comparing the PVAD algorithm and its results from the application to the energy consumption data with the association rule technique and the decision tree technique.

References

[1] N. Ye, T. Y. Fok, X. Wang, J. Collofello, N. Dickson, "Learning partial-value variable relations for system modeling", In Proceedings of the 2018 IEEE ICCAR Conference, Auckland, New Zealand, April 20-23,2018, https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8384628.

[2] Kwok, K green buildings: A comprehensive study of. Y. G., Statz, C., Wade, B., and Chong, W. K., "Carbon emissions modeling for calculation methods", ASCE Proceeding of the ICSDEC (ICSDEC), 118-126, 2012, https://doi.org/10.1061/9780784412688.014.

[3] Swan, L. G., V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques" Renewable and Sustainable Energy Reviews (Sust. Energ. Rev), 13(8), 1819-1835, 2009.

[4] Zhao, H.-X., Magoulès, F., "A review on the prediction of building energy consumption" Renewable and Sustainable Energy Reviews (Renew. Sust. Energ. Rev.), 16(6), 3586-3592, 2012.

[5] PSC Wisconsin, Electricity Use and Production Patterns. Public Service Commission of Wisconsin. Madison, WI: Public Service Commission of Wisconsin. Retrieved from https://psc.wi.gov/thelibrary/publications/electric/electric04.pdf., 2011.

[6] Flex Alert., Peak Demand. Retrieved from Flex Alert Home: http://www.flexalert.org/energy-ca/peak., 2013.

[7] Ahmed, A., et al. Ahmed, A., et al., "Mining building performance data for energy-efficient operation" Advanced Engineering Informatics (Adv. Eng. Inform)., 25(2), 341-354, 2011.

[8] Berges, M. E., et al. "Enhancing Electricity Audits in Residential Buildings with Nonintrusive Load Monitoring", Journal of Industrial Ecology (J. Ind. Ecol.), 14(5), 844-858, 2010.

[9] Fan, C., et al., "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques" Applied Energy (Appl. Energ.), 127(0), 1-10, 2014.

[10] Hawarah, L., et al., User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks, Artificial Intelligence and Soft Computing, Springer Berlin Heidelberg, 6113: 372-379, 2010.

[11] Khan, I., et al., "Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques" Energy Procedia (Enrgy. Proced.), 42, 557-566, 2013.

[12] Khansari, N., et al., "Conceptual Modeling of the Impact of Smart Cities on Household Energy Consumption" Procedia Computer Science (Procedia. Comput. Sci.), 28, 81-86, 2014.

[13] Kim, H., et al., "Analysis of an energy efficient building design through data mining approach", Automation in Construction (Automat. Constr.), 20(1), 37-43, 2011.

[14] Palizban, O., et al., "Microgrids in active network management – part II: System operation, power quality and protection", Renewable and Sustainable Energy Reviews (Renew. Sust. Energ. Rev.), 2014.

[15] Tso, G. K. F. and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks", Energy, 32(9), 1761-1768, 2007.

[16] Vollaro, R. D. L., et al., "An Integrated Approach for an Historical Buildings Energy Analysis in a Smart Cities Perspective" Energy Procedia (Enrgy. Proced.), 45(0), 372-378, 2014.

[17] Friedman, J. H., "Multivariate Adaptive Regression Splines", The Annals of Statistics (Ann. Stat.), 1-67, 1991.

[18] Hastie, T., Tibshirani, R., and Friedman, J. H. The Elements of Statistical Learning, 2nd edition. New York, New York: Springer, 2009.

[19] Zhang, H., and Singer, B. H., Recursive Partitioning and Applications, 2nd edition. New York, New York: Springer, 2010.

[20] Freedman, D. A., Statistical Models: Theory and Practice. Cambridge University Press, 2009.

[21] Bishop, C. M., Pattern Recognition and Machine Learning. Springer, 2006.

[22] James, G., Witten, D., Hastie, T., and Tibshirani, R., An Introduction to Statistical Learning. Springer, 2013.

[23] Ye, N., Data Mining: Theories, Algorithms, and Examples, Boca Raton, Florida: CRC Press, 2013.

[24] Ye, N., The Handbook of Data Mining. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2003.

[25] Gasse, M., Aussem, A., and Elghazel, H., An experimental comparision of hybrid algorithms for Bayesian network structure learning. Lecture Notes in Computer Science: Machine Learning and Knowledge Discovery in Databases, 7523: 58-73., 2012.

[26] Breiman, L., "Random forests" Machine Learning (Mach. Learn.), 45(1), 5–32, 2001

[27] Breiman, L., "Arcing classifier", Ann. Statist (Ann. Stat.), 26(3), 801–849, 1998.

[28] Breiman, L., "Bagging predictors", Machine Learning (Mach. Learn.), 24(2), 123–140, 1996.

[29] Freund, Y., and Schapire, R. E., "Decision-theoretic generalization of on-Line learning and an application to boosting". Journal of Computer and System Sciences (J. Comput. Syst. Sci.), 55(1), 119-139, 1997.

[30] Ho, T. K. "The random subspace method for constructing decision forests", IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE T. Pattern. Anal.), 20(8), 832–844, 1998.

[31] Kam, H. Tim., "Random decision forest.", Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1, 278-282, 1995.

[32] Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. R., "Boosting algorithms as gradient descent" Advances in Neural Information Processing Systems (Adv. Neur. In.), 12, 512-518, 2000.

[33] Schapire, R. E., "The strength of weak learnability" Machine Learning (Mach. Learn.), 5(2), 197–227, 1990.

[34] Jones B. D., Osborne J. W., Paretti M. C., Matusovich H. M., "Relationships among students' perceptions of a first-year engineering design course and their engineering identification, motivational beliefs, course effort, and academic outcomes", International Journal of Engineering Education (Int. J. Eng. Educ.), 30(6), 1340-1356, 2014.

[35] Kline, R. B., Principles and Practice of Structural Equation Modeling, New York, NY: The Guilford Press, 2011

[36] Tsamardinos, I., Brown, L. E., and Aliferis, C. F., "The max-min hill-climbing Bayesian network structure learning algorithm", Machine Learning (Mach. Learn.), 65, 31-78, 2006.

[37] Ellis, B., and Wong, W. H., "Learning causal Bayesian network structures from experimental data" Journal of the American Statistical Association (J. Am. Stat. Assoc.), 103(482), 778-789., 2008.

[38] Akutsu, T., Miyano, S., and Kuhara, S, "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function" Journal of Computational Biology (J. Comput. Biol.), 9(3/4), 331-343, 2000.

[39] Akutsu, T., Miyano, S., and Kuhara, S., "Inferring qualitative relations in genetic networks and metabolic pathways", Bioinformatics, 16(8), 727-734, 2000.

[40] Akutsu, T., Miyano, S., and Kuhara, S., "Identfiication of genetic networks from a small number of gene expression patterns under the Boolean network model", In Proceedings of the Pacific Symposium on Biocomputing, 4, 17-28, 1999.
[41] Bazil, J. N., Qiw, F., Beard, D. A., "A parallel algorithm for reverse engineering of biological networks", Integrative Biology (Integr. Biol.), 3, 1215-1223, 2011.
[42] D'haeseleer, P., Liang, S., and Somogyi, R., "Genetic network inference: From co-expression clustering to reverse engineering" Bioinformatics, 16(8), 707-726, 2000.
[43] Liang, S. "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures", In Proceedings of the Pacific Symposium on Biocomputing, 3, 18-29, 1998.
[44] Marback, D., Mattiussi, C., Floreano, D., "Combining multiple results of a reverse-engineering algorithm: Application to the DREAM five-gene network challenge", Annals of the New York Academy of Sciences (Ann. NY. Acad. Sci.), 1158(1), 102-113, 2009.
[45] Margolin, A. A., Nemenman, I., Basso, K., Wiggin, C., Stolovitzky, G., Favera R. D., and Califano, A., "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", BMC Bioinformatics, 7, Suppl 1, 1-15, 2006.
[46] Soranzo, N., Bianconi, G., and Altafini, C., "Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: Synthetic versus real data", Bioinformatics, 23(13), 1640-1647, 2007.
[47] Stolovitsky, G., and Califano, A. (eds.), Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference. New York, New Yorl: Blackwell Publishing on Behalf of the New York Academy of Sciences., 2007.
[48] Frank, A., and Asuncion, A., UCI machine learning repository. http://archive.ics.uci.edu.ml, Irvine, CA: University of California, School of Information and Computer Science, 2010.
[49] N. Ye, "Analytical techniques for anomaly detection through features, signal-noise separation and partial-value associations", Proceedings of Machine Learning Research, 77, 20-32, 2017. http://proceedings.mlr.press/v71/ye18a/ye18a.pdf.
[50] N. Ye, "The partial-value association discovery algorithm to learn multi-layer structural system models from system data", IEEE Transactions on Systems, Man, and Cybernetics: Systems (IEEE T. SYST. MAN. CYB.: SYST.), 47(12), pp. 3377-3385, 2017.
[51] N. Ye, "A reverse engineering algorithm for mining a causal system model from system data" International Journal of Production Research (Int. J. Prod. Res.), Vol. 55, No. 3, pp. 828-844, 2017. Published online in July 27, 2016. Eprint link: http://dx.doi.org/10.1080/00207543.2016.1213913.
[52] Quinlan, J. R., "Induction of decision trees". Machine learning (Mach. Learn.), 1(1), 81-106, 1986.
[53] Quinlan, J. R., C4. 5: programs for machine learning. Elsevier, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 2, 380-386 (2019)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Effects of The Low-Resolution 3D Video Clip on Cerebrum Blood Flow Dynamics

Toru Tanimura[1], Hiroki Takada[1], Akihiro Sugiura[2], Fumiya Kinoshita[3], Masumi Takada[4,*]

[1]*Department of Human & Artificial Intelligence Systems, Graduate School of Engineering, University of Fukui, Fukui 910-8507, Japan*

[2]*Department of Radiological Technology, Faculty of Health Science, Gifu University of Medical Science, Gifu 501-3892, Japan*

[3]*Department of Electrical and Computer Engineering, Faculty of Engineering, Toyama Prefectural University, Toyama 939-0398 Japan*

[4]*Department of Nursing, Faculty of Nursing, Yokkaichi Nursing and Medical Care University, Mie 512-8043, Japan*

ARTICLE INFO

Article history:
Received:20 December, 2018
Accepted:05 April, 2019
Online: 26 April, 2019

Keywords:
Near-Infrared Imaging (fNIRS),
body sway,
head-mounted display (HMD),
low-resolution,
3D video clip.

ABSTRACT

We have already succeeded in findings that stereoscopic video clips enhance brain activity. We herein increase the knowledge about the influence of 3D and blurred images on the human body, especially on brain activity. From of old, it has been pointed out that the motion sickness is induced by the blurred images. In this study, stabilogram and cerebral blood flow are measured to investigate the process in the brain during the 3D sickness. Activity in the ventral and dorsal streams is enhanced. The most suitable cutoff frequency for viewing the effects of the dorsal stream are estimated between 0.1–0.3 Hz in a Fourier-Shuffle surrogate data analysis of the cerebral blood flow.

1. Introduction

In recent years, various three-dimensional (3D) video display systems have been developed. Generally, they require 3D glasses, but some recent displays can present binocular and multiaspect autostereoscopic images. However, in either case, there are issues: (1) unpleasant symptoms such as headaches, vomiting, and eyestrain, (2) and a lack of ambience and realism. Especially with Japanese 3D televisions, binocular disparity is set to 1º or less; therefore, dynamic movements cannot be fully expressed. The reason for the eyestrain induced by 3D video viewing is not fully known yet; thus, without an appropriate manufacturing standard for 3D videos and their display systems, excessive measures against visually induced motion sickness (VIMS) have been implemented.

With natural vision, accommodation and vergence are consistent. However, it is generally understood that in 3D video viewing, while the lens is accommodated to the position of the screen that displays the image, the eyes converge at the position of the 3D object (a common reason for eyestrain during 3D image viewing). This discrepancy between the accommodation and vergence is considered to be the reason for the eyestrain from 3D viewing and visually induced motion sickness [1–4].

We can easily view stereoscopic video clips in daily life; however, it has been reported that symptoms such as headaches, nausea, dizziness, and vomiting occurred from VIMS [5]. It is said that the abovementioned symptom is induced by the vestibulo-vegetative reflex [6]. The vestibulo-vegetative system is closely accompanied by the vestibular nuclei, which is also involved in body balance. Compensatory movements in this equilibrium function are controlled by the vestibule-spinal reflex and the vestibule-ocular reflex. In these, the outputs from the vestibular nuclei are projected to the antigravitational muscles and the extraocular muscle, respectively. Therefore, body sway is affected by motion sickness, although there are some processes that control the antigravitational muscles through other nervous pathways (see [7]).

According to Patterson [8], if the viewing conditions are sufficiently bright, the depth of field of a target has a mean difference on the order of 1.0 D (diopter), and the

* Masumi Takada, Email: takada@y-nm.ac.jp

accommodation-vergence conflict discussed above is a problem unique to proximity displays. Factors associated with the depth of field are the pupil diameter and resolution, and viewing conditions of images influence the pupil diameter. Most previous studies used a deep depth of field to prevent blurriness, resulting in a measurement environment quite different from that of everyday life.

Moreover, distribution of the vergence fusional stereoscopic limits in stereoscopic images is desired [9]. Eighty-four percent of all subjects could see a stereoscopic image of two planar images with a binocular disparity of 2º. The target under such a condition is a single target without a surrounding image. When there is no other parallax image, an accommodation-vergence process that creates an image from a double image is considered to function as a positive feedback system [10].

With developments such as the miniaturization of diagnostic equipment, brain science is developing rapidly, and a variety of brain activities are being defined [11]. Brain functional imaging using near-infrared spectroscopy (NIRS) has been developed of late years [12]. This technique is known to be a noninvasive measurement of brain activity, which utilizes the property of a hemoglobin which absorbs near infrared light, detecting the time course of the overall reactivity to activation of the cerebral cortex in subjects in a natural state [13–17]. Applying this to surfaces on the brain, we can observe changes in cerebral blood flow (CBF) within a 2–3 cm depth from the scalp, and recorded polygraph as time series data.

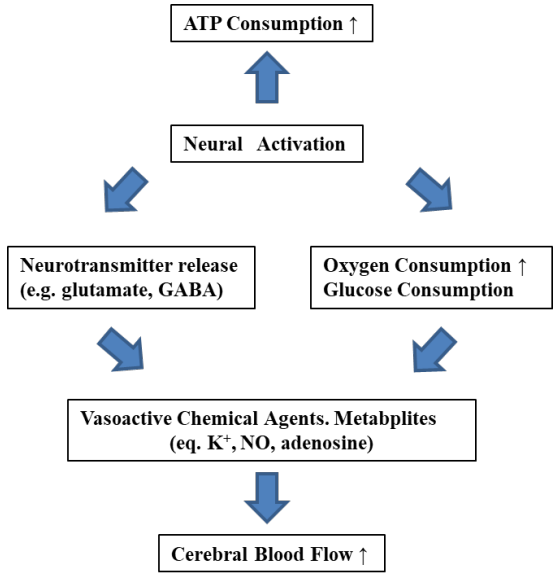


Figure 1: Physiological structure of neurovascular coupling.

The impacts of viewing 3D images on equilibrium function and brain activities were investigated in order to elucidate the cause for 3D sickness, which is induced during the peripheral vision more readily. Potential changes in brain activity can be measured by using fNIRS. Here, peripheral vision for the entire screen was compared to tracking a target while viewing 3D images [18] in accordance with neurovascular coupling (See Figure 1), which was known to be the relationship between subsequent changes in the CBF and local neural activity [19]. Tight temporal and amplitudinal linkages have been observed between the CBF supply and demand of the neuronal activity for over a century [20–22]. According to previous studies [23–25], the regional blood flow is likely controlled by multiple mechanisms such as the feedforward involving neural signaling via neurotransmitters.

In this study, we investigated effects of 3D video clips on human body. In this experiment, we recorded a stabilogram from which the severity of the VIMS could be estimated. The brain function, especially in the hemodynamics in the CBF, was also measured.

2. Material and Method

Biometric data were obtained for the center of pressure (CoP), heart rate variability, and hemodynamics on the surface of the cerebrum in 11 healthy young individuals (mean ± standard deviation: 22.6 ± 1.0 years) with no abnormalities in the extremities and no past medical history of ear or nervous system disease. Moreover, the visual acuity of subjects with the naked eye and/or contact lenses had to be greater than 0.8 and capable of stereoscopic vision. This experiment was fully explained to the subjects beforehand, and written consent was obtained. The experiment was approved by the Ethics Committee of the Department of Human and Artificial Intelligent Systems in the Graduate School of the Engineering University of Fukui (No. H2018010-11).

In this study, subjects wore a head-mounted display (HMD) GOOVIS G2 (Lets-co.jp, Nagoya), and the following video clips (VCs) were projected to the HMD:

VC1) stereoscopic video clip with a resolution of 1080p

VC2) stereoscopic video clip with a resolution of 360p

A resolution of 1080p (1920 × 1080 px; also known as Full HD (high definition video) or FHD (full high definition 2K) and BT.709 (broadcasting service television 709)) is a set of HDTV (high-definition television) high-definition video modes characterized by 1,920 pixels displayed across the screen horizontally and 1,080 pixels down the screen vertically. The *p* stands for progressive scan, *i.e.* noninterlaced [26]. Contents using as visual stimuli in this experiment was supplied by Sky Crystal (Olympus Memory Works Ltd. Co., Tokyo), which was modified with the company's permission. A sphere in a video clip was ambulated in a complex manner.

Biometric data such as the stabilogram, electrocardiogram (ECG), and oxy-/deoxygenated hemoglobin concentrations on the CBF were recorded while the participant viewed high-resolution video clip VC1 and low-resolution video clip VC2. At a sampling frequency of 100 Hz, each sway of the CoP was recorded by using a Wii balance board (Nintendo, Kyoto). Subjects were instructed to maintain the Romberg posture during the duration of the trials. The subjects were asked to use peripheral vision for VC1 for the first 60 s and VC2 for the next 60 s, and to stand when there were no images (resting state). This trial was repeated five times (Figure 2).

In addition to the electrocardiogram (ECG), the CBF on was recorded on the surface of the frontal lobe (1–12 ch), on that of left temporal lobe (13–24 ch), on that of the right temporal lobe (25–36 ch), and on that of the occipital lobe (37–48 ch) in accordance with the fNIRS; channels in the prove of the FOIRE-3000 (Shimadzu, Kyoto) were arranged as shown in Figure 3. Our previous study gave a description of the position where the probe cap was attached in detail [27, 28]. Observing cross-section MRI shots in the brain, we have also confirmed the field on which the CBF changes [29]. Time series data of the CBF were smoothed by the low-pass filtering whose cutoff frequency was set to be 0.15 Hz.

Viewing high-resolution video clip	Viewing low-resolution video clip	Static Upright Posture
60 s	60 s	30 s
Stabilogram, fNIRS, ECG		

× 5 times

Figure 2: Experimental protocol.

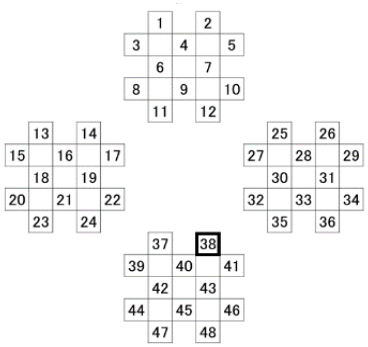


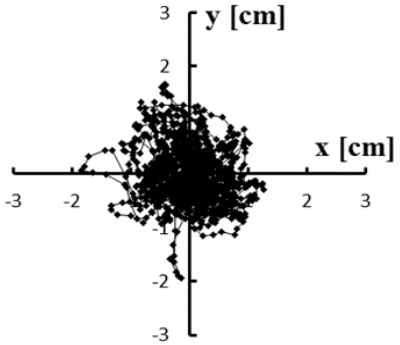
Figure 3: NIRS measurement channel layout.

3. Results

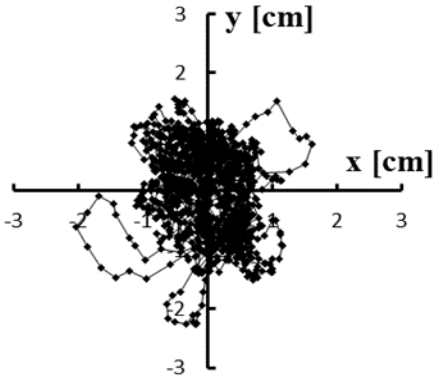
Sway values such as the area of sway, total locus length, total locus length per unit area, and sparse density [30,31] were estimated from the stabilograms (Figures 4–5). By using the Wilcoxon signed-rank test, we compared the sway values while viewing the VCs for a trial as follows. The statistical significance was herein set to be 0.05.

There was no statistical significance in the difference between the sway values while viewing the VCs during the trials except for the total locus (Figure 6b). In the third trial, the total locus length per unit area while viewing VC1 tended to be different from that while viewing VC2 ($p < 0.1$).

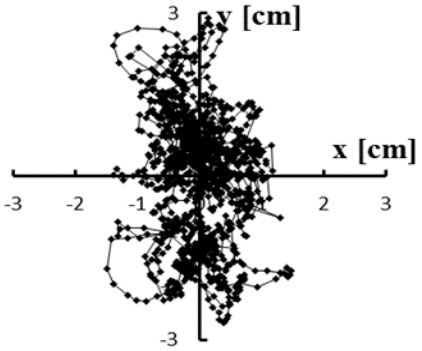
(a)



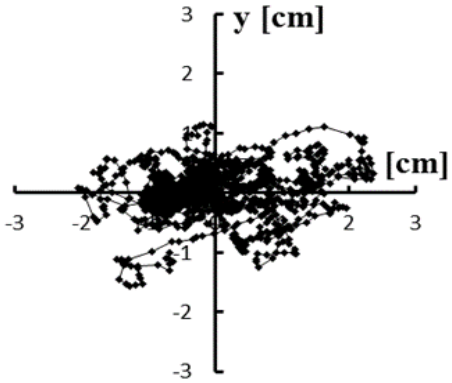
(b)



(c)



(d)



(e)

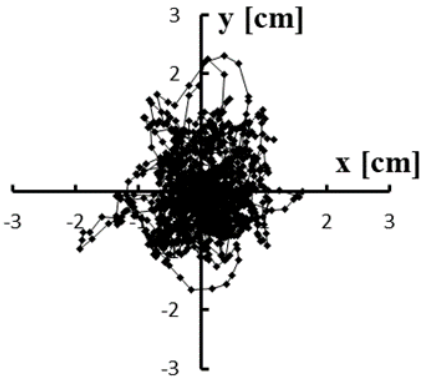


Figure 4: (a–e) Typical stabilograms while viewing VC1 for first to fifth trials, respectively

The oxygenated hemoglobin concentration in the CBF recorded while viewing VC2 was compared to that while viewing VC1. Significant changes were observed in the frontal lobe, temporal lobe, and upper occipital lobe (Figure 7).

For area of sway, sparse density, and total locus length, the sway values for the *n*-th trial (n = 3, 4, 5) were significantly greater than those for the first trial while viewing VC1 ($p < 0.05$). In addition, all sway values for the *n*-th trial (n = 3, 4, 5) were significantly greater than those for the first trial while viewing VC2, as shown in Figure 6 ($p < 0.05$).

4. Discussion

In this study, we investigated effects of 3D video clips on the equilibrium function and the brain function, especially in the hemodynamics in the CBF. Video clips were projected on an HMD.

In this experiment, no statistical differences were observed between the sway values while viewing the high-resolution video clip (VC1) and those while viewing the low-resolution video clip (VC2). The VIMS did not result from differences between the resolutions in this study. In addition, there was a statistical significance between the sway values of the *n*-th trial and those of the first trial (n = 3, 4, 5) while viewing any video clip. The changes in the sway values were considered to be owing to acclimatization in the upright posture.

Time series data are prone to noise with measurement errors, and biological signals are no exception. Noise including in the biological signals should be removed. We have already conducted a study for a nonlinear analysis of the changes in the CBF using fNIRS in order to determine an appropriate cutoff frequency for the low-pass filtering of the CBF. In the experiment, subjects did not view stereoscopic VCs through the HMD but those on a liquid crystal display (LCD) through polarizing spectacles, following a test with subjects' eyes closed (*Cnt*). The VCs were played back for 70 s in visual pursuit and for 70 s in the peripheral vision in succession, which were classified as the resting, the visual pursuit, and the peripheral vision terms. This protocol was repeated five times. Changes in the CBF were recorded in the measurements taken from 10 healthy persons that sat in a dark room. Sampling frequency and the arrangement of the probes were set to be same as shown in this study (Figure 3). Also, the study focused on the dorsal stream, which can be measured around the channel 38. Low-pass filters with some cutoff frequencies were applied to the original CBF for 38-ch, and the translation errors of the smoothed CBF were estimated using the Wayland algorithm [32]. We herein focused on the first set of results from each subject to compare among translation errors during the resting, the visual pursuit, and the peripheral vision terms Because the regularity of the changes in the CBF were indicated through the experiment. Using the Fourier shuffle surrogate data method, the nonlinearity of the results was also analyzed [33].

In each instance, the translation error peaked at cutoff frequencies of 0.3–0.5 Hz [34]. Statistically significant differences were noted between mathematical models of the raw data in the CBF and their surrogate sequences. In addition, statistically significant differences/tendencies were also observed between mathematical models of the surrogate data and the CBF time series smoothed by the low-pass filtering whose cutoff frequencies ≤ 1 Hz. We believed that the CBF data were affected by noises with high frequencies > 1Hz, owing to muscular activities in the extraocular muscles and physical artifacts.

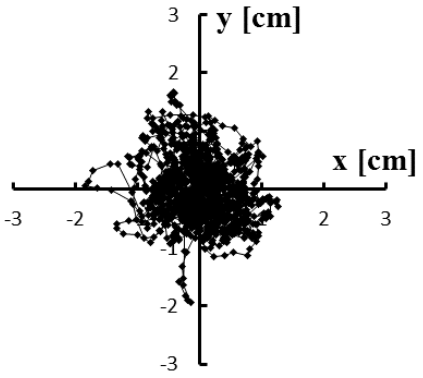
The translation errors were compared according to the cutoff frequencies f_0 (Figure 8). At f_0 = 0.2 Hz, translation errors estimated in the visual pursuit and in the peripheral vision were significantly greater than those observed while the eyes were closed (*Cnt*). Hence, the visual effect is thought to lie in the 0.1–0.3 Hz band.

In order to evaluate the difference between raw data and surrogate data, a *t*-value was herein calculated at f_0 = 0.2 Hz. The values were estimated as 6.77, 3.52, and 1.09 for the *Cnt*, the peripheral vision, and the visual pursuit, respectively. For the visual pursuit and peripheral vision, the *t*-value was reduced compared to the *Cnt*. Thus, it was difficult to observe the nonlinearity in the CBF while viewing stereoscopic video clips. Compared to that for the peripheral vision, the *t*-value was reduced for the visual pursuit. This is thought to be the effect of increased activity in the dorsal stream that governs subjects' spatial awareness with respect to the object being viewed [35]. This manifested itself more strongly when viewing stereoscopic video clips. For this reason, it is thought that a low-pass filter with a cutoff frequency of approximately 0.2 Hz is suitable for observing the nonlinearity and changes in the CBF during the viewing of stereoscopic video clips.

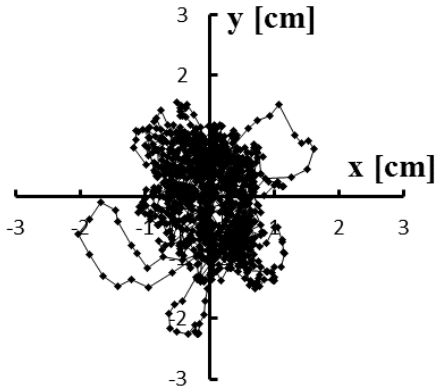
Presently, we investigated effects of 3D video clips on the equilibrium function [27] and the CBF [28,34]. In the former, video clips were projected on the screen. The images were composed of a number of balls. A virtual evoked postural response (VEPR) was expected to be induced by their periodic motion with 0.25 Hz. Severity of the VIMS could be measured as a VEPR. We confirmed that the motion sickness was occurred while viewing 3D video clips. Based on the results of the stabilometry frequency analysis, we observed a reduction in the intensity of power spectral density for 0.25 Hz with the course of the noise amplitudes. The frequency seemed to be synchronized with the periodic motion of the balls in the video clips. In the latter, subjects viewed video clips on LCDs of the video game and the television, respectively. Except for the video game, subjects viewed video clips through shutter/polarizing spectacles in this paragraph.

We investigated the effects of 3D video clips on the brain function, especially in the hemodynamics in the CBF which was remarkably increased while viewing the stereoscopic video clip, compared to while viewing the 2D.

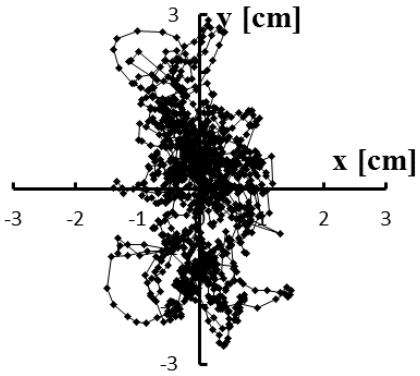
(a)



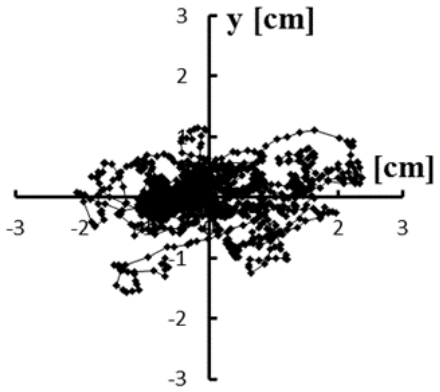
(b)



(c)



(d)



(e)

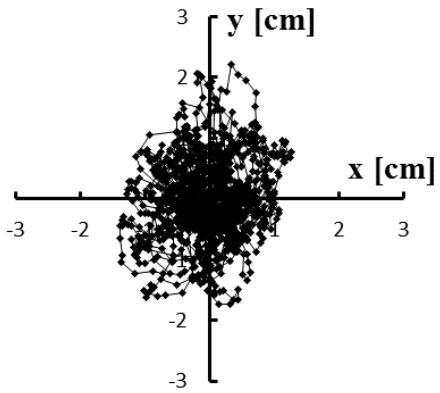
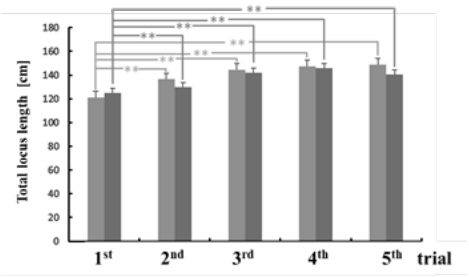


Figure 5: (a–e) Typical stabilograms while viewing VC2 for first to fifth trials, respectively

One question remains: compared with the impact of viewing 3D videos with peripheral vision on the balance system, the amount of sway during viewing is small. Presently, we have not directly captured the instability in the balance system during the peripheral viewing of 3D videos. Therefore, it would be meaningful to improve the measurement precision of the impact on the balance system through numerical models. Considering research ethics, as it would lessen the burden on the subjects, this would not be limited to the design evaluation analysis of ultrahigh-definition images and VR, but could also contribute to hygiene and biomedical engineering.

(a)



(b)

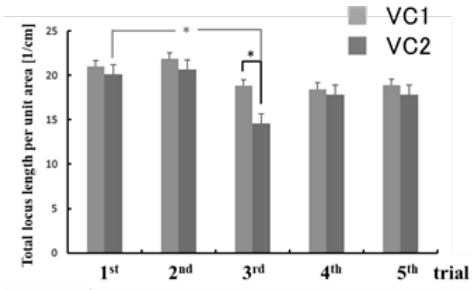


Figure 6: Comparison of sway values while viewing VCs for each trial: total locus length (a); total locus length per unit area (b).

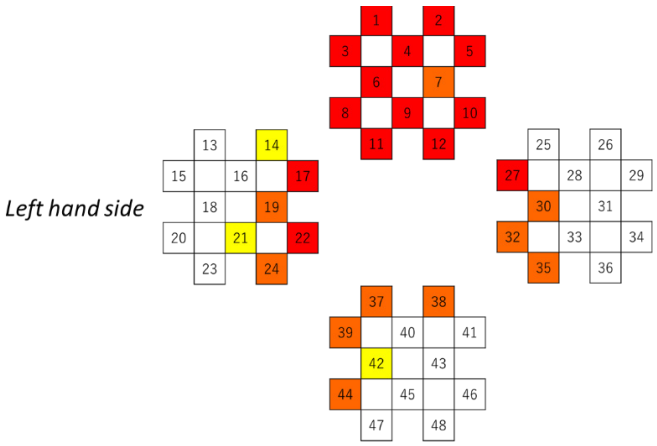


Figure 7: Changes in oxygenated hemoglobin concentration in the CBF recorded while viewing VC2 compared to that while viewing VC1: significant increase ($p < 0.01$ and $p < 0.05$) marked by ■ and ■, respectively; increase tendency ($p < 0.1$) marked by ■.

In this study, by continuing the research challenges described above, we verified the illusion of self-motion (vection) induced by 3D video viewing. As an application, we determined new research topics. Is vection induced by 3D video viewing caused selectively by peripheral vision? Does uncoordinated movement of central and peripheral vision cause vection, inducing visually

induced motion sickness? We plan to elucidate these questions. By working on new research topics, we may be able to propose a guideline for the safe viewing of 3D videos.

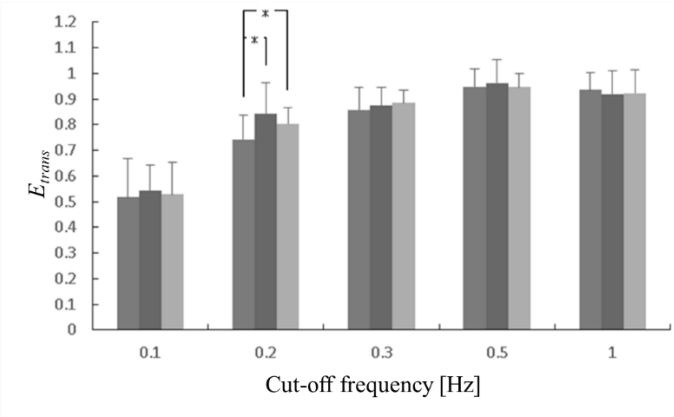


Figure 8: Comparison of translation errors estimated from CBF with eyes closed as control (*Cnt*) ■ with those of the visual pursuit ■, and of the peripheral vision ■.

According to the fNIRS measurement, as shown in Figure 7, significant changes were observed in the temporal lobe and upper occipital lobe that corresponded to the ventral and dorsal streams, respectively [35]. It is difficult to recognize the visual objects and motion processes owing to the low resolution of the video clip. Activity in the ventral and dorsal streams was enhanced, and theircooperativeness might be expected to be found in the next step.

The number of elderly people with dementia in Japan is estimated to reach approximately 7 million people by 2025. It is important to diagnose/respond dementia in early stage. Games for people with mild cognitive impairments (MCIs) has been developed in accordance with the previous. Epidemiologic research [36]; Untreated poor vision is associated with cognitive decline, particularly Alzheimer disease. It is important to take care the decline in the visual acuity. Protecting eyesight also keep our brain function. Their medical evaluations are now conducting, using fNIRS and electroencephalograms (EEG) [37-38]. The basic experiments will be also stated in detail for the virtual space and the video games in the next paper.

5. Conclusion

We focused on the visual consciousness and examined whether the hemodynamics in the CBF depend on the consciousness. Despite the consciousness, the CBF contains body artifacts owing to the cardiovascular system and extraocular muscles in accordance with a nonlinear analysis of the CBF filtered at cutoff frequencies >0.5 Hz. In other words, noise at high frequency (>0.5 Hz) exhibits regularity.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by the Takeuchi Scholarship Foundation and the Descente and Ishimoto Memorial Foundation for the Promotion of Sports Science; the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) Number 17K00715 and Young Scientists (B) Number 15H06711.

References

[1]. F. M. Toates, “Vergence eye movements” Doc. Ophthalmol., 37, 153–214, 1974.
[2]. C. Cruz-Neira, D. J. Sanding, T. A. DeFanti, “Virtual reality: The design and implementation of the CAVE,” in Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’93, 135–142, 1993.
[3]. D. M. Hoffman, A. R. Girshick, K. Akeley, M. S. Banks, “Vergence-accommodation conflicts hinder visual performance and cause visual fatigue” J. Vis., 8(3): 33.1–30, 2018. doi: 10.1167/8.3.33
[4]. T. Shibata, J. Kim, D. M. Hoffman, M. S. Banks, “The zone of comfort: Predicting visual discomfort with stereo displays” J. Vis., 11(8), 1–29, 2011.
[5]. International Standard Organization, “IWA3:2005 Image safety-reducing determinism in a time series” Phys. Rev. Lett., 70, 530–582, 1993.
[6]. K. R. Brizzee, “Mechanics of vomiting: A mini review” Can. J. Physiol. Pharmacol., 68: 221–229, 1990.
[7]. H. Takada, M. Miyao, S. Fateh (eds.), Stereopsis and Hygiene, Springer Nature, 2018.
[8]. R. Patterson, “Human factors of stereo displays: An update” J. SID, 17(12), 987–996, 2009.
[9]. S. Nagata, “Distribution of vergence fusional stereoscopic limits (VFSL) of disparity in a stereoscopic display” TVRSJ, 7, 239–46, 2002 (in Japanese).
[10]. S. Nagata, “The binocular fusion of human vision on stereoscopic displays: Field of view and environments” Ergonomics, 39(11), 1273–1284, 1996.
[11]. A. M. Chiarelli, F. Zappasodi, F. Di Pompeo, A. Merla, “Simultaneous functional near-infrared spectroscopy and electroencephalography for monitoring of human brain activity and oxygenation” Neurophot., 4(4), 041411, 2017.
[12]. M. Fukuda, M. Mikuni, “A study of near-infrared spectroscopy in depression” J. Clinical and Exp. Med., 219, 1057–1062, 2006.
[13]. A. Zardecki, “Multiple scattering corrections to the Beer-Lambert Law” in Proc. SPIE, 103–110, 1983.
[14]. S. Wray, M. Cope, D. T. Delpy, J.S. Wyatt, E. O. Reynolds, “Characterization of the near infrared absorption spectra of cytochrome AA 3 and hemoglobin for the non-invasive monitoring of cerebral oxygenation” Biochemica et Biophysica A., 933, 184–192, 1988.
[15]. T. Kato, A. Kamei, S. Takashima, T. Ozaki, “Human visual cortical function during photic stimulation monitoring by means of near-infrared spectroscopy” J. Cereb. Blood Flow Metab., 13, 516–520, 1993.
[16]. Y. Hoshi, M. Tamura, “Detection of dynamic changes in cerebral oxygenation coupled to neural function during mental work in man” Neurosci. Lett., 150, 5–8, 1993.
[17]. C. E. Elwell, C. E. Cooper, M. Cope, D. T. Delpy, “Performance comparison of several published tissue near-infrared spectroscopy algorithms” Anal. Biochem., 227, 54 –68, 1995.
[18]. M. Takada, Y. Fukui, Y. Matsuura, M. Sato, H. Takada, “Peripheral viewing during exposure to a 2D/3D video clip: Effects on the human body” Env. Health Prev. Med., 20, 79–89, 2015.
[19]. A. A. Phillips, F. H. N. Chan, M. M. Z. Zheng, A. V. Krassioukov, P. N. Ainslie, “Neurovascular coupling in humans: Physiology, methodological advances and clinical implications” J. Cerebral Blood Flow & Metabol., 36, 647–664, 2016.
[20]. F. C. Donders, “Die bewegungen des gehirns und die vera: Nderungen der gefa¨ssfu¨llung der pia mater” Schmid’s Fahrbucher., 69, 16–20, 1851.
[21]. A. Mosso, “Sulla circolazione del cervello dell’uomo” Atti R Accad Lincei., 5, 237–358, 1880.
[22]. C. S. Roy, C. S. Sherrington, “On the regulation of the blood-supply of the brain” J. Physiol., 11, 85–158, 1890.
[23]. D. Attwell, C. Iadecola, “The neural basis of functional brain imaging signals” Trends in Neurosci., 25, 621–625, 2002.
[24]. M. Lauritzen, “Reading vascular changes in brain imaging: Is dendritic calcium the key?” Nature Rev., 6, 77–85, 2005.
[25]. K. Uludag, D. J. Dubowitz, E. J. Yoder, K. Restom, T. T. Liu, R. B. Buxton, “Coupling of cerebral blood flow and oxygen consumption during physiological activation and deactivation measured with fMRI” Neuro Im., 23, 148–155, 2004.
[26]. R. Silva, “720p vs. 1080p: A comparison” Lifewire.com, 2018, last accessed 2016/11/21.
[27]. M. Takada, K. Ito, T. Tanimura, H. Tahara, A. Sugiura, H. Takada, “The impact of 3D games on brain blood flow dynamics and its utilization” in 13th international conference on computer science and education (ICCSE),

828-833, 2018.
[28]. M. Takada, Y. Mori, F. Kinoshita, H. Takada, “Changes in brain blood flow by the use of 2D/3D games” in Universal Accessin Human-Computer Interaction (UAHCI/HCII2016), PartII (eds. A. Margherita, S. Constantin) LNCS9738, 516-523, 2016.
[29]. Y. Matuura, T. Taniguchi, A. Sugiura, M. Miyao, H. Takada, “Distribution of cerebral blood flow during gum-chewing” Forma, 27, 1-4, 2012.
[30]. H. Takada, Y. Kitaoka, S., M. Miyao, “Physical meaning on geometrical index for stabilometry” Equil. Res., 62(3), 168–180, 2003.
[31]. Japan Society for Equilibrium Research (ed.), “Standard of Stabilometry” Equil. Res., 42, 367–369, 1983.
[32]. R. Wayland, D. Bromley, D. Pickett, A. Passamante, “Recognizing determinism in a time series”, Phys. Rev. Lett., 70, 580-582, 1993.
[33]. J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J. D. Farmer, “Testing for nonlinearity in time series: The method of surrogate data” Physica D, 58, 77-94, 1992.
[34]. T. Tanimura, D. Iida, Y. Jono, M. Takada, Y. Matsuura, H. Takada, “Nonlinear analysis using multiple cutoff frequencies for local cerebral blood flow during biofeedback training” JAMMR, 28(6), 1-12, 2018.
[35]. L. G. Ungerleider, M. Mishkin, “Two cortical visual systems” in Analysis of visual behavior (ed. D. J. Ingle et al.), MIT Press, 1982.
[36]. M. A. M. Rogers, K. M. Langa, “Untreated poor vision: A contributing factor to late-life dementia” Am. J. Epidemiol., 171(6), 728–735, 2010.
[37]. G. R. Mendez, I. Dunwell1, E. A. M. Mirón, M. D. V. Cerdán, S. de Freitas, F. Liarokapis, A. R. G. Gaona, “Assessing NeuroSky’s usability to detect attention levels in an assessment exercise” in Human-Computer Interaction (UAHCI/HCII 2009), Part I (ed. J. A. Jacko), LNCS 5610, 149–158, 2009.
[38]. S. J. Johnstone, S. Roodenrys, R. Blackman, E. Johnston, K. Loveday, S. Mantz, M. F. Barratt, “Neurocognitive training for children with and without AD/HD” ADHD Atten. Def. Hyp. Disord., 4, 11–23, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 380-390 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Using Fuzzy PD Controllers for Soft Motions in a Car-like Robot

Paolo Mercorelli *

Institute of Product and Process Innovation, Leuphana University of Lueneburg, Volgershall 1, D-21339 Lueneburg, Germany

ARTICLE INFO

Article history:
Received: 26 September, 2018
Accepted: 24 November, 2018
Online: 05 December, 2018

Keywords:
Fuzzy control
Nonholonomic car-like robot
Sensorless cascade control system

ABSTRACT

This paper deals with the control problem for nonholonomic wheeled mobile robots moving on the plane, and in particular, the use of a Fuzzy controller technique for achieving a given motion task which consists of following a rectilinear trajectory until an obstacle occurs on the path. After a background part, in which the fundamental knowledge of Fuzzy control is considered, the problem of the avoidance of an obstacle is taken into consideration. When an obstacle occurs on the path, the drive assistant provides for its avoidance calculating the minimal distance from which the avoidance maneuver starts. Conditions on the parameters of a PD controller are calculated using a Fuzzy based approach. An observer is designed to obtain unmeasurable states to be used in the control loop. In the Appendix of this paper a formal demonstration of a Proposition is proven in which the convergence of the system state estimation of the observer is shown. Simulations considering a real transporter vehicle for a storage service are shown.

1. Introduction

In the field of logistics, the demand for shorter lead times, the highest possible flexibility and high efficiency increases every year. In order to implement these requirements, it is of a great importance that the resources which are responsible for the flow of materials should be optimized and designed so that they can be intuitive and easy to be used. For transportation today as it was in the past, trucks and pallet trucks are used and it is essential to employ personnel for picking within the warehouse. Since a high potential for improving the picking process is still present, a high level of attention will be paid to the optimization of these tasks.

This paper is an extension of work originally presented in *ICCC conference 2018,* [1]. The difference between this work and that already published in [1] is a wide extension of the simulated results together with an extension of the background aspects related to the Fuzzy control and the Luenberger observer. Moreover, the demonstration of the convergence of the estimation is reported in the Appendix of this paper. In this sense, the problem of the observer is considered in depth because represents one of the most challenging problems. The idea is to control the vehicle using the measurement of the steering angular position and the position of the vehicle. The aim of this work is to create a simulation by which the evasive behavior of a newly developed Picking Vehicle can be tested. The results of this work can be applied at a later stage to a second stage of the vehicle. The simulation allows to test the vehicle's behavior and it gives ideas for the design of the controller parameters. More in depth, this paper deals with the control problem for nonholonomic wheeled mobile robots moving on the plane, and in particular, the problem of the avoidance of obstacles is taken into consideration. When an obstacle occurs on the path, then the drive assistant provides for its avoidance calculating the minimal distance from which the avoidance maneuver starts. Using a Fuzzy approach, conditions on the parameters of a PD controller are calculated. A Luenberger observer is used in the control loop to minimize the number of the sensors. It is known that Luenberger observer is one of the most used observers in motion control and it is a high gain observer, see [2, 3] and [4] and Kalman Filters, see [5] and [6]. Even though the model of a mobile robot is well known in terms of its structure, for controlling mechanisms Fuzzy logic is often used, see [7]. The use of Fuzzy controller is not limited to the system without physical insights but it is shown in [8] that any system can take advantage from such a kind of control strategy. In particular, in [9] the authors describe the design and the implementation of a trajectory tracking controller using Fuzzy logic for mobile robots to navigate in the indoor environments. Most of the previous works used two independent controllers for navigation and avoiding obstacles. Also in [10] the salient feature of the proposed approach is that it combines the Fuzzy gain scheduling method and a Fuzzy proportional-integral-derivative (PID) controller to solve the nonlinear control problem. The paper is organized in the

*Corresponding Author Paolo Mercorelli, Institute of Product and Process Innovation, Leuphana University of Lueneburg, Volgershall 1, D-21339 Lueneburg, Germany, Email: mercorelli@uni.leuphana.de

following way. Session 2 presents the model of a general nonholonomic system. Session 3 is devoted to the presentation of the Luenberger observer. Session 4 is dedicated to the Fuzzy and PD control strategy. Session 5 shows the simulated results and the conclusions close the paper. At the end of the paper in an Appendix, a Preposition is proven in which the convergence of the estimation of the system state variables is shown.

2. The Nonholonomic System

The first step in developing the kinematic model, which will be used below, is the consideration of nonholonomic conditions. The model used is that of robots with car-like properties. Therefore, the properties of the robot will be described by means of the position and the orientation of the vehicle along the route and the angle of the wheels to be controlled. The main feature of the mobile vehicle similar to a robot is the presence of nonholonomic constraints, since they are based on the condition of getting around without slipping on the ground. This means that the movements of the vehicle are restricted and the vehicle cannot move freely in any direction, see [11]. One example is the parallel parking a car. To get into the parking space it is not possible to drive the vehicle easily sideways. It must be moved forward or backward and be maneuvered by means of changing the steering angle of the front wheels into the parking space. To understand the conditions of the nonholonomic system, at the beginning the case of each wheel will be considered. The speed of the wheel is orientated in the direction in which the wheel moves.

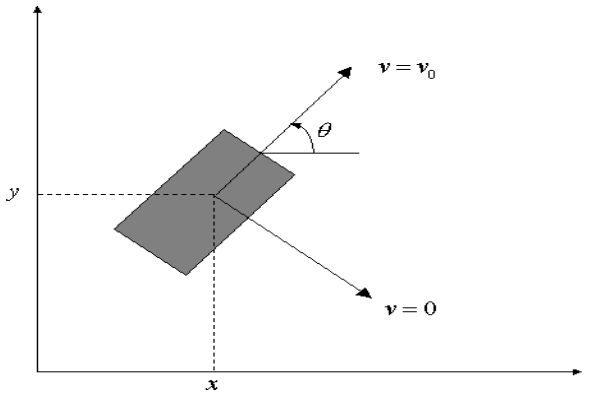


Figure 1. Nonholonomic system: (Mellodge, Feedback Control for a Path Following Robotic Car, 2002)

Due to the nonholonomic properties speed in other directions is not possible. The properties of each single wheel can be described using a vector consisting of three generalized coordinates. The position coordinates x, y within a fixed coordinate system in which the wheel touches the ground and the angle θ, which is the alignment of the wheel to the x-axis. The generalized velocities q cannot accept independent values, they are subject to the following constraint, see Figure 1 from [12]:

$$[\sin\theta \; -\cos\theta \; 0]\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\theta}\end{bmatrix}=0. \tag{1}$$

Equation (1) constitutes a particular expression of a kinematic constraint, a Pfaffian constraint. In this case, the Pfaffian $C(q)\dot{q}=0$, for example, is linearity in the generalized speeds. As a result, all the generalized velocities in the core of the matrix $C(q)$ are included, see [13].

For the individual wheel, this results in the following model:

$$\dot{q}=\begin{bmatrix}\cos\theta\\ \sin\theta\\ 0\end{bmatrix}v_1+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}v_2. \tag{2}$$

In this case, v_1 represents the linear speed of the wheel and v_2 represents the angular velocity around the vertical axis.

2.1. Kinematic Model of the Vehicle with Global Coordinates

The kinematic model of a vehicle is often used because of its simplicity and its accuracy when the vehicle behavior must be shown under normal driving conditions, see [14]. The model of a car-like robot vehicle is an option for the kinematic model. This can be described (see Figure 2) on the basis of four variables, see [14].

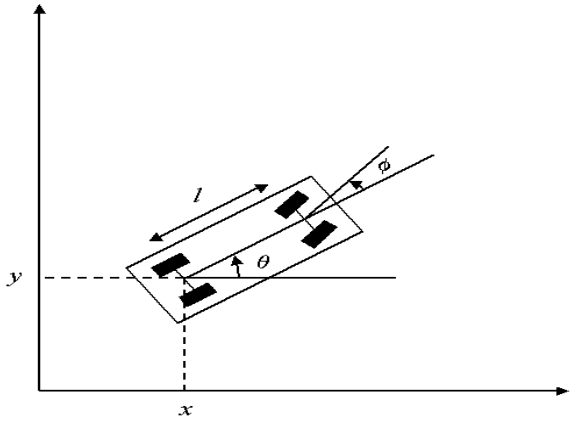


Figure 2. Car model with global coordinates (Source: (Mellodge & Kachroo, 2008), page 31)

x and y represent the Cartesian coordinates of the rear axle of the vehicle. The angle θ is the orientation of the vehicle with respect to the x-axis and the angle ∅ which shows the steering angle of the front wheels of the vehicle, θ with respect to the alignment. Because of front and rear axles, the system is subject to two different nonholonomic conditions:

$$\dot{x}_f\sin(\theta+\emptyset)-\dot{y}_f\cos(\theta+\emptyset)=0$$

$$\dot{x}\sin\theta-\dot{y}\cos\theta=0, \tag{3}$$

where x_f and y_f represent the coordinates of the front axle, x and y represent the coordinates of the rear one. From the conditions shown in Figure 2, the current velocities in x and y directions can be calculated, in which v_1 represents the speed at the rear wheels:

$$\dot{x}=v_1\cos\theta$$

$$\dot{y}=v_1\sin\theta. \tag{4}$$

Taking the rear axle as a reference point and taking into account the distance between the two axes, the result for the front axle is:

$$x_f = x + l\cos\theta$$

$$y_f = y + l\sin\theta. \quad (5)$$

This results in the position of the derivative to the speeds:

$$\dot{x}_f = \dot{x} + l\,\dot{\theta}\cos\theta$$

$$\dot{y}_f = \dot{y} + l\,\dot{\theta}\sin\theta. \quad (6)$$

If the new conditions of formula (2.6) in the nonholonomic conditions of the front axle (2.3) are used and solved for θ, then the constraint of the front axle results to be:

$$\dot{\theta} = \frac{\tan\emptyset}{l} v_1. \quad (7)$$

From this, the entire model results as follows, see also [15]:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\theta}\\ \dot{\emptyset}\end{bmatrix} = \begin{bmatrix}\cos\theta\\ \sin\theta\\ \frac{\tan\emptyset}{l}\\ 0\end{bmatrix} v_1 + \begin{bmatrix}0\\0\\0\\1\end{bmatrix} v_2, \quad (8)$$

v_1 thereby represents the speed of the rear wheels and v_2 represents the speed of the steering angle or the change of the steering wheels. In order to investigate the nonlinear system from the previous equation, the principle of the Lie algebra provides the ability to perform a controllability analysis. The existing system is known to be a nonlinear and a drift-free one due to the nonholonomic conditions of the system. A general description of the kinematic nonlinear system is shown below in equation (9):

$$\dot{q} = G(q)v. \quad (9)$$

Here, q represents the n vector of the general coordinates.

$\boldsymbol{v}$ is the m-vector of the input speeds, and it must be smaller than the amount of the general coordinates (m <n). The matrix G is divided into a plurality of uniform g_i columns, where (i = 1, ..., m) is valid. For the model of the vehicle like robot which is used, n = 4 different coordinates and m = 2 controllable inputs are valid. This results in the following model which is used for the consideration of controllability, see [13].

$$\dot{q} = g_1(q)v_1 + g_2(q)v_2, g_1 = \begin{bmatrix}\cos\theta\\ \sin\theta\\ \frac{\tan\emptyset}{l}\\ 0\end{bmatrix}, g_2 = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}. \quad (10)$$

The two vector fields and g_1 and g_2 allow driving and steering of the model. Apparently only the directions given by g_1 and g_1 are possible, but in reality, it is known that practically any position can be approached with a car. To describe the process mathematics provides the method of Lie derivative, which creates a new vector field out of two vector fields. In [12] the Lie brackets are designed as follows:

$$\Big[g_1, g_2, [g_1, g_2], \big[g_1[g_1, g_2]\big]\Big]. \quad (11)$$

The brackets g_1 and g_2 are already known, the other two arise by means of the Jacobian matrix:

$$J_f(x) = \begin{pmatrix}\frac{df_1}{dx_1} & \frac{df_1}{dx_2} & \cdots & \frac{df_1}{dx_n}\\ \frac{df_2}{dx_1} & \frac{df_2}{dx_2} & \ddots & \vdots\\ \frac{df_3}{dx_1} & \frac{df_3}{dx_2} & \cdots & \frac{df_m}{dx_n}\end{pmatrix} \quad (12)$$

to:

$$[g_1, g_2](x) = \frac{dg_2}{d_x} g_1(x) - \frac{dg_1}{d_x} g_2(x) = \begin{bmatrix}0\\0\\\frac{-1}{l\cos^2\emptyset}\\0\end{bmatrix}, \quad (13)$$

and

$$\big[g_1[g_1, g_2]\big] = \frac{d[g_1,g_2]}{d_x} g_1(x) - \frac{dg_1}{d_x}[g_1, g_2](x) = \begin{bmatrix}\frac{-\sin\theta}{l\cos^2\emptyset}\\ \frac{\cos\theta}{l\cos^2\emptyset}\\0\\0\end{bmatrix}. \quad (14)$$

Further matrices are not present and therefore, the individual matrices can be merged to the overall matrix G. Finally, the rank of the matrix will be determined. Matrix G is the following:

$$G = \begin{bmatrix}\cos\theta & 0 & 0 & -\frac{\sin\theta}{l\cos^2\theta}\\ \sin\theta & 0 & 0 & \frac{\cos\theta}{l\cos^2\theta}\\ \frac{\tan\theta}{l} & 0 & -\frac{1}{l\cos^2\theta} & 0\\ 0 & 1 & 0 & 0\end{bmatrix} \quad (15)$$

From the consideration of matrix G, the rank of this matrix results to be equal to four, because all four columns are independent for all θ. From the middle two columns it is directly visible that they cannot be zero. The two outer columns should be specifically considered, because of the danger that for ∅ = 0 a dependency can occur. For this, the determinant of the matrix G was calculated by the symbolic toolbox of MATLAB. This investigation showed that except in the case of a higher or lower steering angle of ∅ = ± π/2 there is no dependency. It follows that the system is controllable, as long as it is within the valid steering angle, see [5].

2.2. Kinematic Model of the Vehicle with Path Coordinates

The Global Coordinates model of the vehicle presented in subsection (3.1) is useful for the purpose of simulation and of good use to gain first insights into the behavior. However, for an application in practice, it is not particularly suitable in most cases because the sensors have difficulty to determine the position of the vehicle in relation to global coordinates. Therefore, a new model must be introduced for the rest of the work. This describes the position of the vehicle as dependent of the path (Figure 3). The distance, which the vehicle must keep to the wall and to obstacles

is indicated by the variable d and measured by the laser scanner, which is located centrally on the vehicle front side. The distance between the front and the rear axle is defined by l. The angle θ reproduces the orientation of the vehicle with respect to the x-axis again as also in the global coordinate system. θ_t indicates the angle between a tangent to the path and the x-axis. In order the vehicle to be able to follow the path, an angle θ_p which indicates the difference between the orientation of the vehicle and the angle of the tangent θ_t is defined:

$$\theta_p = \theta - \theta_t. \tag{16}$$

For the desired path, tracking this difference should tend as much as possible to zero. The distance traveled is defined by s and is an arbitrarily selectable starting position. The course of the track, or for instance, the curvature is given by c (s) and defined as follows:

$$c(s) = \frac{d\theta_t}{ds} \; or \; \dot{\theta}_t = \; c(s)\dot{s}. \tag{17}$$

Furthermore, the speeds along the track $\dot{s}$ and the speed required perpendicular to the distance $\dot{d}$ for the model will be required:

$$\dot{s} = v_1 \cos\theta_p + \dot{\theta}_t d$$

$$\dot{d} = v_1 \sin\theta_p. \tag{18}$$

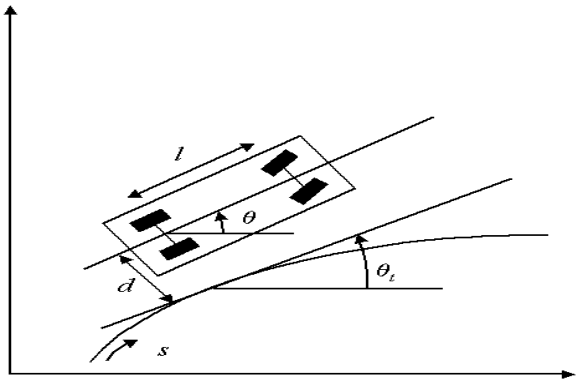


Figure 3. Kinematic Model with Path coordinates (Source: (Mellodge & Kachroo, 2008), page 33)

The overall model based on the redefined variables with Path Coordinates, as it can be seen in Equation (19), can be set up.

$$\begin{bmatrix} \dot{s} \\ \dot{d} \\ \dot{\theta}_p \\ \dot{\emptyset} \end{bmatrix} = \begin{bmatrix} \frac{\cos\theta_p}{1-dc(s)} \\ \sin\theta_p \\ \frac{\tan\emptyset}{l} - \frac{c(s)\cos\theta_p}{1-dc(s)} \\ 0 \end{bmatrix} v_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} v_2. \tag{19}$$

3. Observer

The state variables must be supplied to the controller for the control of dynamic systems or control of the physical parameters of the track. These quantities are usually measured. However, if this is not possible, or only possible with a great effort, they have to be determined in another way. For this, an observer can be used, which reconstructs a state (condition) based on the course of input and output variables. Here, the model of the controlled system is connected in parallel to the actual process and supplied with the same input dimension. If the model is correct, it performs the same action as the control path and differences can thus only occur through the different initial conditions. If the control system is stable and it is possible to wait long enough, the observer and the controlled system commute into the same sizes and it is: $\hat{y}(t) = y(t)$, and $\hat{x}(t) = x(t)$ and $\hat{x}(t)$ can be used as a state feedback. This simple variant has major lacks:

- The control path $x(t)$ will only consider a disturbance size, while the model $x(t)$ remains unchanged, when a disturbance occurs. Therefore, this method is only suitable for undisturbed, uninteresting cases for control.
- Secondly, a difference $x(t) - \hat{x}(t)$ generated by different initial states $\hat{x}_0$ and x_0 is only compensated when the controlled system is stable. Out of the equation of the state of the controlled system:

$$\dot{x}(t) = Ax(t) + Bu(t), x(0) = x_0 \tag{20}$$

with $\hat{x}(t)$ the expression follows:

$$x(t) - \hat{x}(t) = e^{At}(x_0 - \hat{x}_0). \tag{21}$$

This however is only possible for a stable matrix A against zero, regardless of whether the unstable controlled system is stabilized within the control loop by a state feedback. Since only the system matrix of the controlled system and not of the closed circuit enters into the relationship, the simple observer is useful only for stable systems. Based on this principle, David G. Luenberger has developed an extension of this model in 1964. However, this observer is expanded to the difference between the output of the model y and the output of the control path y. This difference is used in order to equalize the state of the model to the route. This method has already proven itself in the control loop, there, however, the deviation of the controlled variable is used to perform a control intervention, which influences the controlled variable. Based on the following non-jump-capable control system the subsequent design of the observer follows:

$$\dot{x}(t) = \mathrm{Ax(t)} + \mathrm{B}u(\mathrm{t}),$$

$$x(0) = x_0$$

$$y(t) = Cx(t). \tag{22}$$

In the design of the observer, the system model is supplemented by the input u_B:

$$\dot{\hat{x}}(t) = \mathrm{A}\hat{x}(\mathrm{t}) + \mathrm{B}u(\mathrm{t}) + u_B(\mathrm{t}), \hat{x}(0) = \hat{x}_0$$

$$\hat{y}(t) = C\hat{x}(t). \tag{23}$$

In the new input $u_B$the difference between the measured output of the control path y and the output of the model $\hat{\boldsymbol{y}}$ is returned to be:

$$u_B(t) = L\left(y(t) - \hat{y}(t)\right). \tag{24}$$

If equations (3.4) and (3.5) are put together, we obtain the following relationship:

$$\dot{\hat{x}}(t) = \mathrm{A}\hat{x}(\mathrm{t}) + \mathrm{B}u(\mathrm{t}) + LC\big(x(t) - \hat{x}(t)\big). \tag{25}$$

After merging, the effect of the newly introduced input is recognized. As long as $\hat{x}(t) = x(t)$ is true, the right part falls away, because the addend becomes zero and the model thus runs in parallel to the control path. If a deviation between the input variables, which influence the outputs, occurs, so the behavior of the model due to the feedback will be influenced. In order to reduce the deviation, a matrix L must be found which decreases the difference of $x(t) - \hat{x}(t)$. The input vector $x(t)$ is not known in practice as a rule, however, the output vector $y(t)$, serving as an input of the observer. The result for the observer is a controlled dynamic system with the two inputs u(t) and y(t) as in equation (3.7).

$$\dot{\hat{x}}(t) = (\mathrm{A} - LC)\hat{\mathrm{x}}(\mathrm{t}) + \mathrm{B}u(\mathrm{t}) + Ly(t). \tag{26}$$

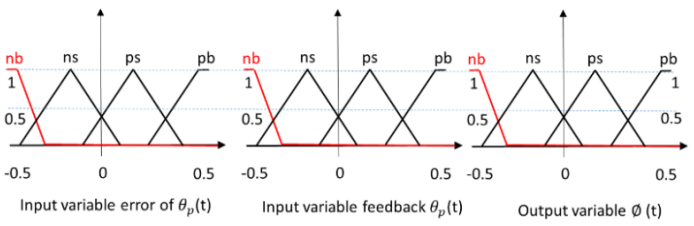
In the simulation, the observer has the following characteristics:

$$\begin{bmatrix} \dot{\hat{s}} \\ \dot{\hat{d}} \\ \dot{\hat{\theta}}_p \\ \dot{\hat{\emptyset}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \hat{s} \\ \hat{d} \\ \hat{\theta}_p \\ \hat{\emptyset} \end{bmatrix} + \begin{bmatrix} \frac{\cos\theta_p}{1-dc(s)} \\ \sin\theta_p \\ \frac{\tan\emptyset}{l} - \frac{c(s)\cos\theta_p}{1-dc(s)} \\ 0 \end{bmatrix} v_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} v_2 + \begin{bmatrix} k1 \\ k2 \\ k3 \\ k4 \end{bmatrix} \big[d(t) - \hat{d}(t)\big]. \tag{27}$$

The principle of the Luenberger observer is based on a linear system, in which matrices with fixed values are integrated. However, the system in the work is a nonlinear one, in which the matrix B represents a field with variables. The principle of the observer remains the same as the one in the linear version. The stability of the observer seems to be given, however, it was only heuristically determined. To investigate the stability more precisely, the stability theory of Lyapunov is used and in the Appendix at the end of the chapter a formal analysis is conducted.

4. Fuzzy Control Algorithm integrated with a PD controller

In addition to methods for mathematically analytical problem solving of control and monitoring tasks, now human experience and knowledge that cannot be expressed mathematically for problem solving are being increasingly used. For this purpose, the Fuzzy logic is often used. In contrast to the classical control engineering, the behavior of a system is tried to be described with traditional means and then a controller will be tried to be designed with analytical methods. Fuzzy systems are particularly well suitable for the design of systems under a vague knowledge. For example, they are well suitable for not exactly known process or a not exactly known controller behavior. Due to the completely different approaches for both controller designs, completely different solution approaches for Fuzzy and classical approaches for regulation are used. While in the classical, for instance PD scheme, in a first step a model of the controlled-track is formed, in the second step the associated controller design follows only using the formed path. This procedure can therefore be described as a model-driven interpretation. The Fuzzy controller represents a contrast for it, since, this is controller-oriented. Subsequently, the basic concepts of Fuzzy control, and the steps of the design of a Fuzzy controller are described. For execution of the control mechanism in Fuzzy logic three actions must be performed. In the first part, the sharp values, which are supplied by the input variable, will be transformed into the fuzzification. The physical values of input values become the converted Fuzzy sets. These are a specific form of Fuzzy sets, in which the values are not described as usually in mathematics by numerical variables in the form of sharp numbers, but by colloquial expressions. Here formulations arise such as: when the distance to the wall is too small, the distance must be increased somewhat by increasing the steering angle. Here, the descriptions “distance slightly too small” and “steering angle slightly increase” represent the blurry relationship between the distance to the wall and the steering angle. The variables, which result from this, are called linguistic variables and they will be assigned to a membership function. The membership function makes an indication, in what proportion between 0 and 1, the value of a blurry statement is true. It also provides a statement to what extent an element belongs to a certain set. In literature fuzzy controller is used also for nonholonomic systems. In [16] a fuzzy PD controller for a dynamic model of nonholonomic mobile manipulator in order to treat the trajectory tracking control and to eliminate the effect of external force on the end-effector is proposed. Concerning the robustness of the Fuzzy approach, the paper in [17] addresses the output feedback trajectory tracking problem for a nonholonomic wheeled mobile robot in the presence of parameter uncertainties, external disturbances, and a lack of velocity measurements. A combination between a heuristic Fuzzy and PID controller is also designed in [18] to move the robot upward or downward the inclined plane and approach the target point. In [19] fuzzy adaptive observers together with parameter adaptation laws are designed to estimate the state-dependent disturbances in both kinematics and dynamics in order to adapt a Fuzzy adaptive controller.



In [20] a holistic, for holonomic and nonholonomic systems, intelligent control strategy is proposed in which only position measurements are used.

In Figure 4, the two inputs and the output of one of the two Fuzzy controllers are presented as an example. Here the different

Fuzzy sets of the two inputs and the output can be seen. The linguistic variable "ns" means "negative small" and is associated with an interval 0.1 to -0.4. Within the interval there are different degrees of association, the value 0.15 with a factor of 1 is the most strongly associated value and the association decreases in negative and positive direction until at 0.1 and -0.4 the association of 0 is achieved. Such linguistic variables are defined over the entire area to be controlled, if possible, they must have an overlap at their adjacent variable, so that no areas remain undefined. In order to provide a high degree of flexibility in different executions such as trapezoidal, Gaussian, or triangular shaped, the associated functions can be defined and combined with each other. MATLAB provides for different execution forms. In the present work, the trapezoidal and triangle-shaped execution have been chosen.

Table 1. Fuzzy rules

Output		Angle			
		nb	ns	ps	pb
Error	nb			nb	nb
	ns			ns	
	ps	ps	ps		
	pb	pb			

Output		Distance			
		nb	ns	ps	pb
Error	nb			nb	nb
	ns			ns	ns
	ps	ps	ps	ps	ps
	pb	pb	pb	pb	pb

In Table 1 all the used fuzzy roles in this application are indicated. For instance, the linguistic variable "nb" stands for "negative big," "ps" stands for "positive small" and "pb" stands for positive big.

After the fuzzification was performed, the inference follows. The rules for evaluation will be developed with help of an expert or by a private experience. These consist of two parts, the if-condition (premise): IF the distance is too small and the Fuzzy inference (conclusion): THEN the steering angle must be increased slightly. The input variables will be linked by means of an AND-operation in the IF part. Then the resulting rules are connected by means of an OR-operation. So, the following rules resulted for the conditions used in this work. The possible Fuzzy sets for the deviation from the desired value are referred here as an error. The Fuzzy sets that are delivered as a response of the system, are referred to as a feedback. A 4x4 matrix follows from it, in which not every field is filled. The actual inference takes place within three steps after the rules having been defined:

1) Aggregation (evaluation of the rule premises): The rule premises are evaluated here, that is the IF-parts of rules corresponding to their affiliations and selected operators. The input variables were associated with an AND-operator, in this case a min-operator results to be used, which always selects the smallest affiliation value of a Fuzzy set. From a human point of view, it acts pessimistically and not compensatory.
2) Implication (evaluation of the conclusions (THEN-part)): Evaluation of the conditions occurs during the implication. The rules can also only be partially true in the Fuzzy control, in contrast to the classical control. Therefore, in this step, a definition follows in which the dimension of the condition is true or not. The minimum procedure will be used for it (see Figure 5), which assumes as a conclusion the minimum of the premise (IF-part) and conclusion (THEN- part). In simple words, this means that the output function as a result of this operation, is cut off at the level of fulfillment of the premise and an area in the output Fuzzy set arises.
3) Accumulation (summary of all rules); The conclusions of all the rules will be summarized in the accumulation. This is necessary because, as a rule not only one rule is true, but several rules can be at least partially valid. Due to the fact that the rules are connected with each other through the Or-links, here the maximum operator is used in the selection of the conclusions. Here the maximum value of the Fuzzy sets is used as the output function, see Figure 6. Since the Min-method was used for the implication and the Max-method was used for the accumulation, the process which is applied in the work is called a max / min inference.

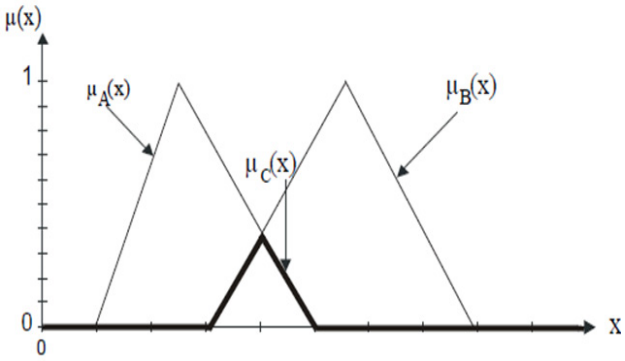


Figure 5. Minimum Operator

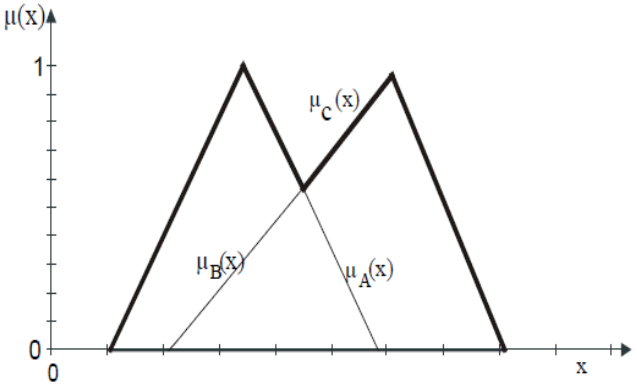


Figure 6. Maximum Operator

Defuzzification is carried out during the last step, after the inference was completed. A new Fuzzy set will turn out to be from the inference, and thus a Fuzzy information will be delivered. So that other participants can process the signals, sharp output values must be generated again. This is done using the gravity method. Here, a Fuzzy set resulting from the inference is considered as a totality.

As shown in Figure 7, the Fuzzy set became a set B^* which determines the output variable. The centroid is now determined by the gravity method and the gravity coordinate u_s gives it as a

defuzzification result. The coordinate of the centroid is calculated by the following formula:

u_s= sharp output value

u_i= abscissa bases

$\mu_B * (u_i)$= membership degree for u_i

q =number of abscissa bases

$$u_s = \frac{\int_{u_{min}}^{u_{max}} u*\mu_B*u(d_u)}{\int_{u_{min}}^{u_{max}} \mu_B*u(d_u)} \approx \frac{\sum_{i=1}^{q} u_i*\mu_B(u_i)}{\sum_{i=1}^{q} \mu_B(u_i)} \quad (28)$$

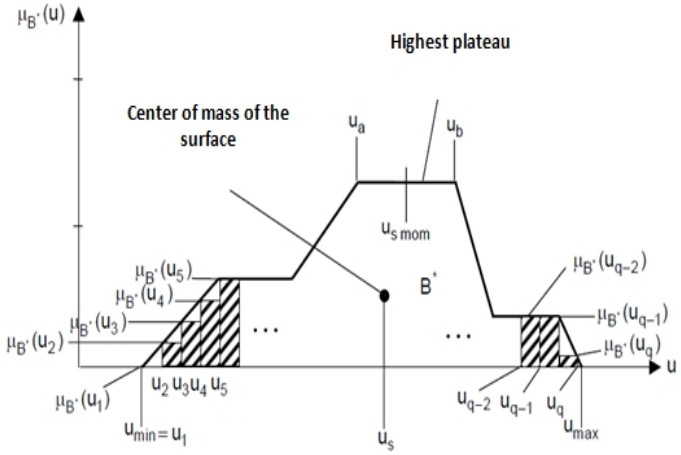


Figure 7. Defuzzification by the centroid method (centroid)

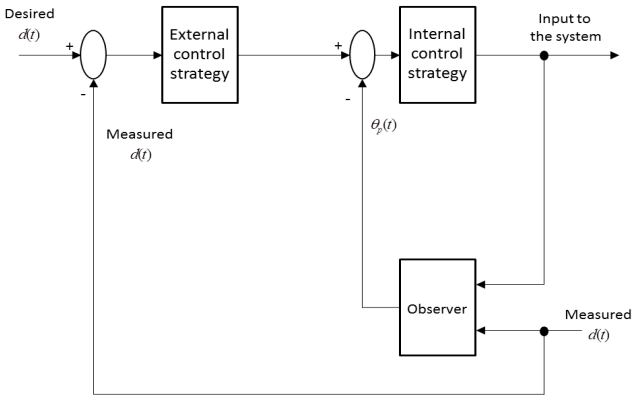


Figure 8. Simplified diagram of the cascaded control loop

During the calculation according to the gravity method, all active rules determine the sharp output. The calculation methods and criteria for the selection of operators are selected in Simulink within the Fuzzy Editor and run subsequently automatically.

Figure 8 shows the general structure of the cascade control scheme including the Luenberger observer. In Figure 8 it is visible how the observer is issued. Through the measurement of the distance between the car and the wall all other state variables can be observed.

Here, the outer and the inner circle of the PD and the Fuzzy control with help of an adder were interconnected. This is shown in Figure 9.

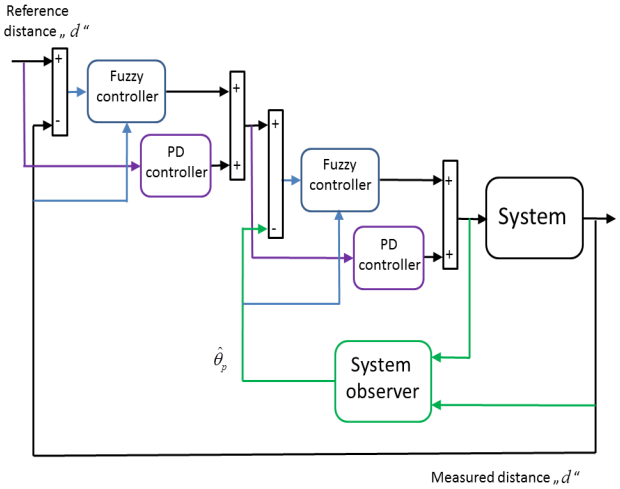


Figure 9. Simulink diagram combining PD and Fuzzy controllers in a cascade control structure together with an observer

Remark 1

The choice of this particular Fuzzy logic is due to the fact that, considering more variables than proposed, not relevant improvements are obtained in terms of reduction of error.

4.1. A Gaussian Curve to Avoid Obstacle

A Gaussian function offers the possibility to obtain the highest possible, incremental and without jumps resulting function to avoid obstacles. This guarantees due to its properties, in addition to a slow and continuous adjustment of the steering angle, a high degree of stability in its derivation. This property proves to be of a great advantage, because the derivative is required for further calculations within the simulation. To achieve a high degree of flexibility and influence a formed function, which is similar to a Gaussian function, but offers more free parameters, and thus becomes a Gaussian-like function. The calculation is based on the freely selectable amplitude H, which determines the maximum level of the function. In order to influence the abdomen of a function, upsetting or a stretching factor was introduced. This will be automatically calculated based on the obstacle width in a sub-function. The following formula is a result for the curve:

H = amplitude of the curve

o = factor for upsetting or stretching of the curve

t = velocity factor

$$G(t) = H * e^{-\left(\frac{t}{o}\right)^2} \quad (29)$$

The derivative of the curve is thus obtained as follows:

$$\dot{G}(t) = -H * e^{\left(\frac{t}{o}\right)^2} * 2\left(\frac{t}{o}\right) * \frac{1}{o} \quad (30)$$

Factor “t” states, in this case, the speed of the vehicle and the phase shift, which serves as a correction factor. In the following part the overall configurations of the various control concepts and their implementation in Simulink were presented.

5. Results

Figure 10 shows that the cascade PD control structure increases significantly faster the distance to the wall than the other two control configurations. This requires only 8 seconds, in the meantime the cascade Fuzzy control structure, represented in Figure 11, requires about 12 seconds to bring the vehicle at the same distance. The combination of both controllers manages to do it in just over 10 seconds, see Figure 12. During the fallback procedure, which starts at about 25 seconds, the quality of the results of the controller changes, the PD controller shows here his weakness. In the case of the example shown, this operation supplies during the evacuation process the worst results. It swings at the desired distance from ≈∓ 6 cm. The Fuzzy Control System in comparison provides results of ≈∓ 5 cm, which accounts for a difference of ≈17%. The combination of the two controllers presents the best results. This results in a maximum deviation of ≈∓ 2 cm, which means a difference of ≈67% compared to the PD controller. Thus, the combination of the two controllers in the simulation provides the best results, although it increases the distance to the wall something slower than the PD it performs, but it has a significantly higher stability during the fallback procedure. It is important to be able to configure the smallest possible protective field because the aisles in which the vehicle should move anyway already offer little space and thus the smallest possible protective field is necessary. A comparison between the model of the kinematic system and the estimated values of the observer based on the distance d is also visible.

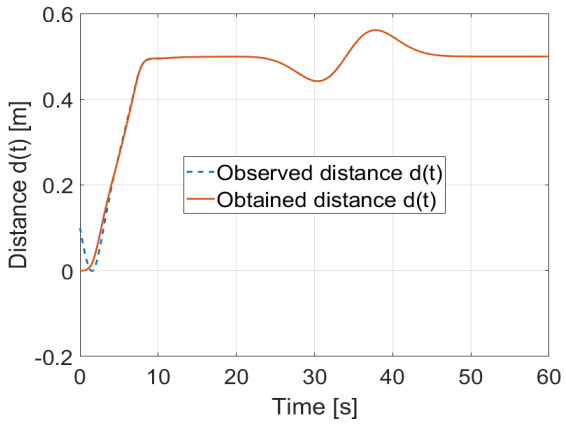


Figure 10. Distance of the car from the wall using the cascade PD control structure

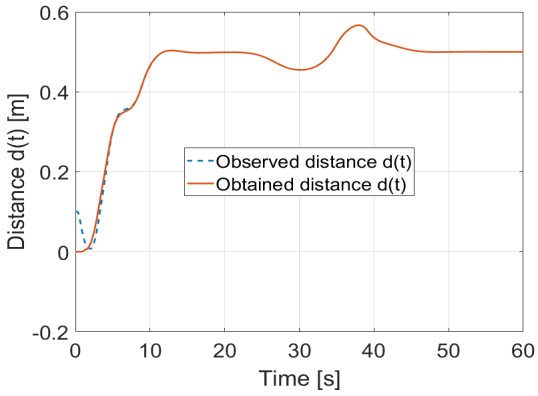


Figure 11. Distance of the car from the wall using the cascade Fuzzy control structure

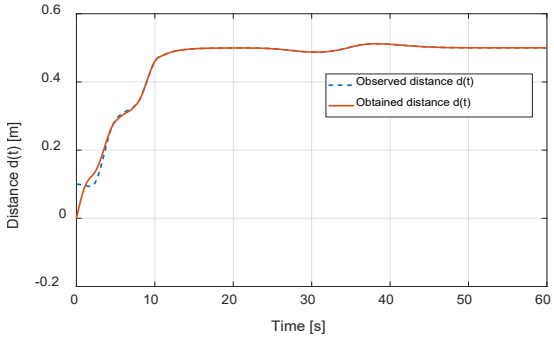


Figure 12. Distance of the car from the wall using the combination between PD and Fuzzy cascade control structure

To ensure that the vehicle does not slip during the evacuation process even with a heavy load, or does not even lose some of the load and also to be able to configure a minimum possible protective field, the steering angle may not exceed a maximum impact of ∓ 10°. Figures 13, 14 and 15 show the time progress of the steering angle in degrees using PD, Fuzzy and a combination of PD and Fuzzy controllers respectively. As it can be seen there, the requirement of the maximum steering angle has been met for all three schemes.

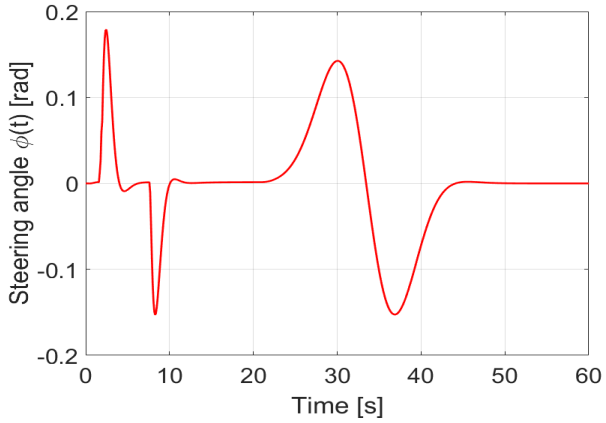


Figure 13. Steering angle Ø using the PD cascade controller structure

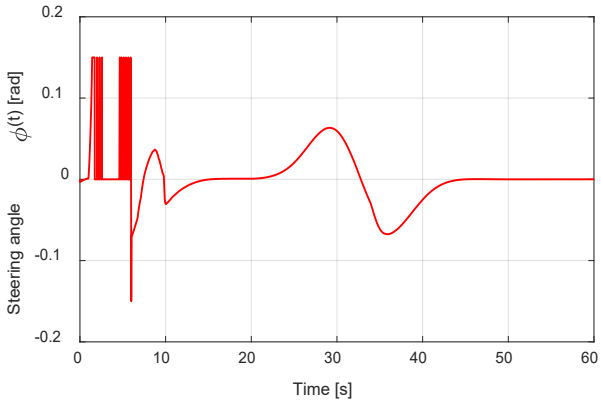


Figure 14. Steering angle Ø using the Fuzzy cascade controller structure

However, it was shown that the PD control with ∓ 10 °, during the evasive maneuver, is still within the required range, but the other two controls are significantly lower. Thus, during the evasive maneuver, the Fuzzy control with a range of ≈∓8 ° and

the combination of both controls, PD and Fuzzy controllers, lead the comparison with a range of ≈∓6 °. As it has already been described, also the combination of the two controls leads if we compare the simulation results, because it requires the least amount of the steering angle. Advantages of this are, as it has already been explained above, that no abrupt changes of direction as a result of a strong steering maneuver occur due to the small steering angle and the avoidance maneuver has fewer risks like slippage of the vehicle, or a shifting or falling down of the load. Figures 13, 14 and 15 show the simulation results concerning these aspects.

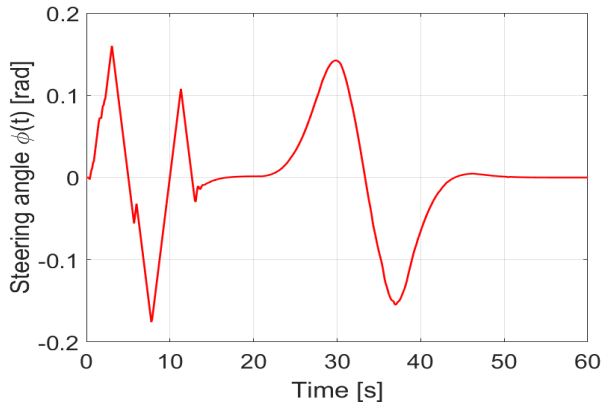


Figure 15. Steering angle ∅ using the combination between PD and Fuzzy cascade control structure

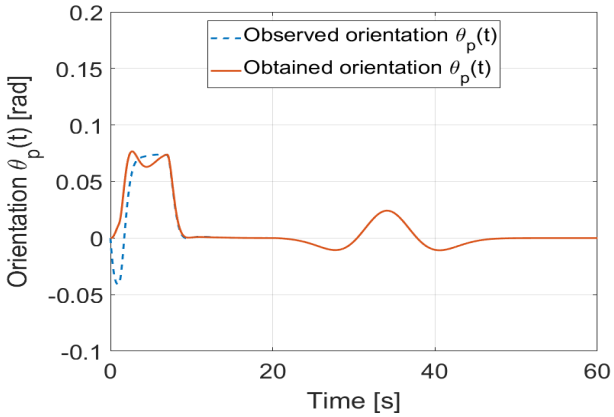


Figure 16. Angle θ_p using the PD cascade controller structure

5.1. Results of the Vehicle Orientation

The vehicle as it is described must follow the predetermined path. For this purpose, the differential angle $\theta_p$was introduced which describes the difference between the angle to the path θ_t and the actual vehicle heading θ. Figures 16, 17 and 18 show the time course of the difference angle θ_p expressed in degrees. For the route to follow this best, the difference angle must be kept as small as possible. The initially large deflection is due to the correction, in order to achieve the already mentioned predetermined distance and in this case, less relevant, since this is not to be taken into account for the actual viewing of the path. Interestingly, the course is between 20 and 50 seconds since the avoidance of the obstacle occurs, which goes along with the path. Here, PD and Fuzzy controllers deliver almost equivalent results, but the maximum excursions hardly differ so they both achieve maximum angle of ≈∓1.5° during the evasive maneuver. It is different with the result of the combination of the two schemes, this is as well as the distance and the steering angle significantly better. Here is a maximum angle of ≈∓ 0.5° during the evasive maneuver and thus the best result. A comparison between the model of the kinematic system and the estimated values of the observer based on the angle θ_p is also visible.

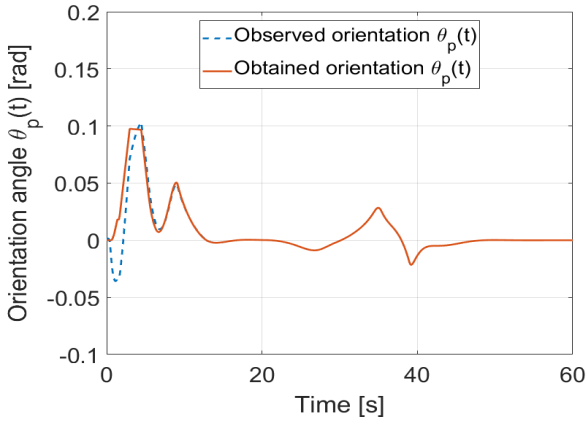


Figure 17. Angle θ_p using the Fuzzy cascade controller structure

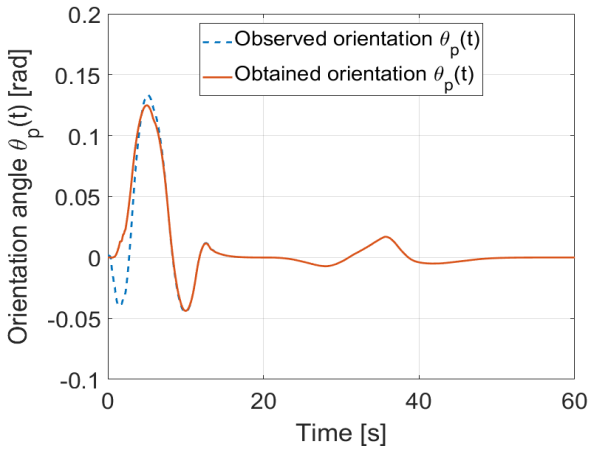


Figure 18. Angle θ_p using the combination between PD and Fuzzy cascade control structure

Figure 19 shows the error of the distance and Figure 20 shows the error of the orientation of the car. In both cases it is possible to see how after 45 seconds the error results to be almost equal to zero.

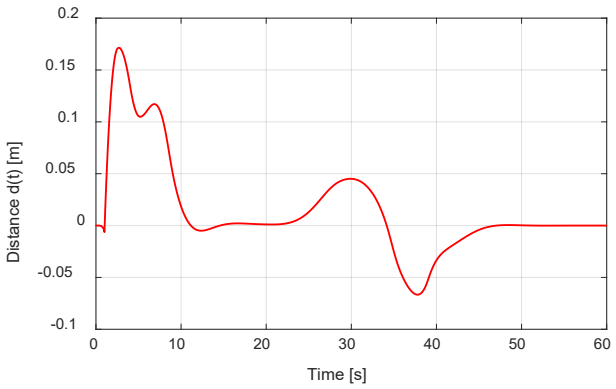


Figure 19. Error of the distance of the car from the wall

Remark 2

To discuss the obtained results with the already existing contributions it is possible to say that in this contribution an original combination of Fuzzy control strategy combined in a PD controller in a cascade structure is presented. This idea is an original one, which is not present in the already existing literature as already discussed through the cited literature.

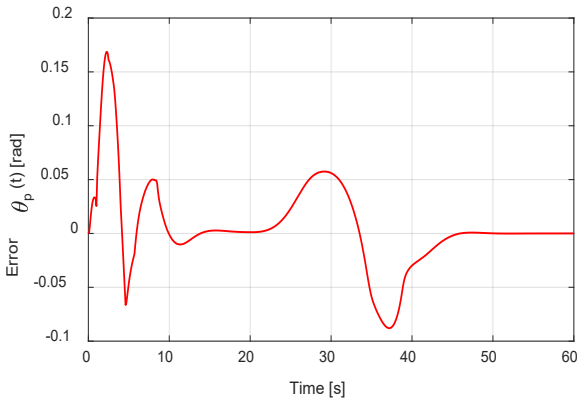


Figure 20. Error of angle θ_p

6. Conclusions

This paper deals with the control problem for nonholonomic wheeled mobile robots moving on the plane to avoid obstacles. Parameters of a PD controller are calculated using a fuzzy based approach. To estimate the orientation of the vehicle a Luenberger observer is involved in the control scheme. In the context of the Luenberger observer, the demonstration of the convergence of the estimation of the proposed observer is shown in the Appendix of the paper. Simulations considering a real transporter vehicle for a storage service are shown.

References

[1] P. Mercorelli, "Fuzzy based control of a nonholonomic car-like robot for driveassistant systems", *2018 19th International Carpathian Control Conference (ICCC)*, Szilvasvarad, pp. 434-439, 2018.

[2] P. Mercorelli, "An adaptive and optimized switching observer for sensorless control of an electromagnetic valve actuator in camless internal combustion engines". *Asian Journal of Control*, 16:959–973, 2014.

[3] P. Mercorelli, "A two-stage sliding-mode high-gain observer to reduce uncertainties and disturbances effects for sensorless control in automotive applications". *IEEE Transactions on Industrial Electronics,* 62(9):5929–5940, 2015.

[4] P. Mercorelli, "A motion-sensorless control for intake valves in combustion engines". *IEEE Transactions on Industrial Electronics,* 64(4):3402–3412, *2017*.

[5] P. Mercorelli, "A two-stage augmented extended Kalman filter as an observer for sensorless valve control in camless internal combustion engines". *IEEE Transactions on Industrial Electronics,* 59(11):4236–4247, 2012.

[6] P. Mercorelli, "A hysteresis hybrid extended Kalman filter as an observer for sensorless valve control in camless internal combustion engines". *IEEE Transactions on Industry Applications,* 48(6):1940–1949, 2012.

[7] C. Zheng, Y. Su, P. Mercorelli, "A simple fuzzy controller for robot manipulators with bounded inputs". *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM,* 737–1742, *2017*.

[8] G. Feng, "A Survey on Analysis and Design of Model-Based Fuzzy Control Systems". *IEEE Transactions on Fuzzy Systems,* 14: 676-697, 2006. doi: 10.1109/TFUZZ. 2006.883415.

[9] H. Omrane, M. S. Masmoudi, M. Masmoudi, "Fuzzy Logic Based Control for Autonomous Mobile Robot Navigation", *Computational Intelligence and Neuroscience,* 2016, Article ID 9548482.

[10] Y. L. Sun, M. J. Er, "Hybrid fuzzy control of robotics systems", *IEEE Transactions on Fuzzy Systems,* 12: 755-765, 2004. doi: 10.1109/TFUZZ.2004.836097.

[11] F. Fahimi, *Autnomous Robots Modeling, Path Planning, and Control.* New York: Springer Science & Business Media, 2009.

[12] P. Mellodge, "*Feedback Control for a Path Following Robotic Car". M.S. Thesis,* Blacksburg: VirginiaTech, 2002.

[13] J.-P. Laumond, *Robot Motion Planning and Control.* London: Springer, 1998.

[14] P. Mellodge, K. Pushkin, *Model Abstraction in Dynamical Systems: Application to Mobile Robot Control.* Berlin Heidelberg: Springer, 2008.

[15] A. De Luca, G. Oriolo, C. Samson, "Feedback control of a nonholonomic car-like". In Jean-Paul Laumond, editor, *Robot Motion Planning and Control,* 4: 171–253. Berlin Heidelberg: Springer, 1998.

[16] A. Karray and M. Feki, "Tracking control of a mobile manipulator with fuzzy PD controller", *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, Hammamet, pp. 1-5, 2015.

[17] S. Peng, W. Shi, "Adaptive Fuzzy Output Feedback Control of a Nonholonomic Wheeled Mobile Robot", in *IEEE Access*, vol. 6, pp. 43414-43424, 2018.

[18] M. Roozegar, M. J. Mahjoob, "Modelling and control of a non-holonomic pendulum-driven spherical robot moving on an inclined plane: simulation and experimental results", in *IET Control Theory & Applications*, vol. 11, no. 4, pp. 541-549, 24 2, 2017.

[19] D. Chwa, "Fuzzy Adaptive Tracking Control of Wheeled Mobile Robots With State-Dependent Kinematic and Dynamic Disturbances", in *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 3, pp. 587-593, 2012.

[20] Y-C. Chang, B-S. Chen, "Intelligent robust tracking controls for holonomic and nonholonomic mechanical systems using only position measurements", in *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 491-507, 2005.

[21] G. Zhu, A. Kaddouri, L. A. Dessaint, O. Akhrif. "A nonlinear state observer for the sensorless control of a permanent-magnet ac machine". IEEE Transactions on Industrial Electronics, 48(6):1098–1108, 2001.

Appendix

Stability analysis of the Luenberger observer

To analyze the convergence of the estimation of the proposed Luenberger observer, the well-known Lyapunov approach is adopted, [4], [21]. In the following part the stability analysis is proposed.

Proposition 1

Let us consider the continuous observer defined by (3.8), if the nonlinear functions present in this system are Lipschitz ones, then there exists a matrix

$$\mathrm{K} = \begin{bmatrix} k1 \\ k2 \\ k3 \\ k4 \end{bmatrix}$$

such that

where the corresponding estimation error dynamics are given by:

$$\dot{e}(t) = (\mathrm{A} - \mathrm{KC})e(\mathrm{t}) + \Delta f(\mathrm{t}), \qquad \text{(A1)}$$

with

$$\hat{e}(\mathrm{t}) = \begin{bmatrix} \hat{s} \\ \hat{d} \\ \hat{\theta}_p \\ \emptyset \end{bmatrix} - \begin{bmatrix} \hat{s} \\ \hat{d} \\ \hat{\theta}_p \\ \hat{\emptyset} \end{bmatrix}$$

and with A_0 = A−KC

where

$$\Delta f(\mathrm{t}) = \begin{bmatrix} \frac{\cos\theta_p}{1-dc(s)} \\ \sin\theta_p \\ \frac{\tan\emptyset}{l} - \frac{c(s)\cos\theta_p}{1-dc(s)} \\ 0 \end{bmatrix} v_1 - \begin{bmatrix} \frac{\cos\hat{\theta}_p}{1-\hat{d}c(s)} \\ \sin\hat{\theta}_p \\ \frac{\tan\emptyset}{l} - \frac{c(s)\cos\hat{\theta}_p}{1-dc(s)} \\ 0 \end{bmatrix} v_1,$$

in which term Δf states the uncertainty due to the not perfect cancellation between the dynamics of the system and the dynamics of the observer.

Proof 1: With (A,C) being an observable pair, matrix A_0 for a suitable choice of the observer gain K is a Hurwitz matrix. This means that there exist symmetric and positive matrices P_0 and Q_0 which satisfy the so called Lyapunov equation

$$A_0^T P_0 + P_0 \, \mathrm{A}_0 = -Q_0. \qquad \text{(A2)}$$

To show the asymptotic stability of (A1), this Lyapunov function is introduced:

$$\mathrm{V}\,(\mathrm{e(t)}) = \frac{\mathrm{e}^T(t)P_0e(t)}{2}.$$

The time derivative is given by:

$$\dot{\mathrm{V}}\left(\mathrm{e(t)}\right) = \dot{\mathrm{e}}^T(t)P_0e(t) + \mathrm{e}^T(t)P_0\dot{e}(t).$$

From (A1) it follows

$$\dot{\mathrm{V}}\left(\mathrm{e(t)}\right) = (\mathrm{A}_0e(t) + \Delta f)^T P_0e(t) + \mathrm{e}^T(t)P_0(\mathrm{A}_0e(t) + \Delta f)P_0.$$

This yields:

$$\dot{\mathrm{V}}\left(\mathrm{e(t)}\right) = (\mathrm{e}^T(t)A_0^T + \Delta f^T)P_0e(t) + \mathrm{e}^T(t)P_0(\mathrm{A}_0e(t) + \Delta f)P_0$$

At the end considering (A2), it follows:

$$\dot{\mathrm{V}}\left(\mathrm{e(t)}\right) = -\mathrm{e}^T(t)Q_0e(t) + \Delta f^T P_0e(t) + \mathrm{e}^T(t)P_0\Delta f.$$

Thus Δf is a Lipschitz function, then there is a positive constant L such as

$$||\Delta f(x_1, x_2)|| \leq L||x_1, x_2||.$$

If ρ_1 is the small eigenvalue of matrix Q_0 and ρ_2 largest eigenvalue of matrix P_0, if the following conditions result satisfied:

$$\rho_1 \geq 2\rho_2,$$

then

$$\dot{\mathrm{V}}\left(\mathrm{e(t)}\right) \leq -(\rho_1 - 2\rho_2)e^2(t).$$

Once suitable matrices Q_0 and P_0 are chosen, we can also choose matrix K such that A_0 has negative real eigenvalues to guarantee (A2).

ASTES

Special issue on Recent Advances in Engineering Systems Journal Vol. 3, No. 5, 421-426 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Effect of Various Parameters for Temperature Distribution in Human Body: An Analytic Approach

Kabita Luitel[*,2], Dil Bahadur Gurung[1], Kedar Nath Uprety[2]

[1]*Department of Mathematics, School of Science, Kathmandu University, 45210, Nepal*

[2]*Central Department of Mathematics, Tribhuvan University, 44618, Nepal*

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 08 October, 2018
Online: 18 October, 2018

Keywords:
Heat transfer,
Temperature profiles
Modified Bessel's Equation

ABSTRACT

This paper is the extension of the work originally presented in 2nd International Conference on Man and Machine Interfacing (MAMI 2017) and the paper of Yue et al. The present study aims at observing the effect of various parameters on temperature distribution profiles at various environmental temperature, tissue thermal conductivities, metabolic rates, blood perfusion rates, and heat transfer coefficients. The analytic solution of Pennes' bioheat equation in the steady-state case is obtained by using the Modified Bessel's equation incorporating the effect of sweating and non-sweating state of the body. From the study, it was observed that the variation of atmospheric temperature and heat transfer coefficients have a significant effect for the temperature distribution in the body towards the skin surface.

1. Introduction

Heat transfer plays a very important role in the living system. This is the complex process, which includes not only some physical factors—conduction, convection, radiation, evaporation but also the physiological factors—blood flow and metabolism. These physical and physiological factors help maintain the constant human body temperature around 37^0C which is the equilibrium point of the production of heat and loss of heat by the body. The extreme temperatures from the normal body temperature influence the function of biological tissue and the whole system of the body. Death may occur if the body temperature is 27^0C and below and if 42^0C and above. So we should maintain the body temperature around 37^0C. The heat transfer in the blood vessel also helps to maintain uniform body core temperature regardless of changes in environmental temperature. Metabolism, another important source of heat gain is the chemical process that occurs in the living organism to grow and produce, maintain their structure, and respond to their respective environment [1, 2, 3].

Sweat evaporation is one of the effective parameters in the thermoregulatory process, which is the only way to lose heat when the ambient temperature is higher than the normal body (36.1^0C - 37.2^0C) temperature. Due to the sweat evaporation, 22% of heat losses from the body. Evaporative heat exchange also involves the loss of heat through the evaporation of sweat from the skin surface. We generally calculate the rate of sweat evaporation as the weighted mean value of the body core and skin temperature. The reasonable equation for the sweat rate is [4, 5].

$$E = 8.47 \times 10^{-5}\{(0.1 \times T_{sk} + 0.9 \times T_b) - 36.6°\text{C}\} Kg/m^2/sec$$

Where T_{sk} = skin surface temperature

and T_a = 37°C (body core temperature)

It has been proved that the transfer of latent heat from the living being to its environment is often estimated by multiplying the loss in weight attributable to evaporation by the latent heat vaporization of water 'L' which decreases from 2501J/g at 0^0C to 2406 J/g at 40^0C[6]. Havenith et al.[7] however, have mentioned the values for 'L' ranging from 2,398 J/g to 2,595J/ g. and finally suggested the latent heat of evaporation is only dependent on temperature giving a number of 2,430J/g at 30^0C.

Acharya et al. [5] used variational finite element (FEM) method to prepare one dimensional heat transfer model for the comparative study of temperature profiles of human male and

[*]Corresponding author: Kabita Luitel, Department of Mathematics, Bhaktapur Multiple Campus, TU, Nepal, E-mail: kabi123luitel@gmail.com

temperature profiles of female luteal and follicular phases of the menstrual cycle.

Havenith et al. [7] studied experimentally by using thermal manikin to determine the effective cooling power of moisture evaporation. They measured both heat loss and mass loss independently by allowing a direct calculation of an effective latent heat of evaporation.

The knowledge of heat transfer is equally important in the field of biomedical research as well as in the treatment of cancer, now a great threat to the existence of humanity. Bioheat transfer models have therapeutic and clinical importance. These models are helpful for the more effective treatment of cancer than it is now. As per Arkin et al., the effect of hyperthermia treatment depends on the temperature and duration of heating. If a constant temperature could be maintained, the duration of heating would be a reasonable way of expressing thermal dose [8].

Over a hundred years' time, the effects of the blood flow have been examined on the heat transfer of living tissue. After Bernard came up with an experimental study in 1876, physicians, physiologists, and engineers are interested in the mathematical modeling of the complex thermal interaction between the vascular systems of the body.

In 1948 pennes' H.H. [9] proposed the simple linear mathematical model based on experimental observation for describing heat flow within the tissue. Many other researchers, one after another, have developed alternative models for describing the perfusion rate and difference between the arterial blood temperature and the local tissue temperature. But Pennes' model still has an acceptable result to predict the transient temperature due to its simplicity and flexibility [1].

Previous researchers such as Acharya et al. [5], Khandey and Saxena [10], Gurung and Saxena [11], Aijar and Dar [12], Nadel [13] observed the effect of latent heat of sweat evaporation for temperature variation in the human body only by using numerical (Specially FEM) techniques. Numerical methods give the approximation result whereas if the analytic solutions of these equations are attainable, they will give the exact result. Even though the analytical method was used by Yue et al.[2] to solve the model, but this model has not incorporated the effect of Latent heat and sweat evaporation. The analytical approach for the study of such problems is still lacking. So the present study focuses on a mathematical model of the body temperature based on the Pennes' bioheat equation due to the sweat evaporation. The temperature at various physical and physiological conditions in the sweating case in comparison with the non-sweating case will be observed in the study

2. Mathematical Model

The governing differential equation used in the model is given by [9].

$$\rho c \frac{\partial T}{\partial t} = \nabla.(K\nabla T) + M(T_a - T) + S \qquad (1)$$

Where, $\boldsymbol{\rho}$ = density of tissue (kg/m^3), $\boldsymbol{c}$ = specific heat (J/Kg. ^{0}C), $\boldsymbol{K}$ = thermal conductivity (W/m^{0}C)**,** ρ_b = density of blood (Kg/m^3), $\boldsymbol{w_b}$ = blood perfusion rate per unit volume (m^3/s.m^3), $\boldsymbol{c_b}$ = blood specific heat (J/kg.^{0}C), M = $\rho_b w_b c_b$ (W/m^3.^{0}C), $\boldsymbol{S}$ = metabolic heat generation (W/m^3), T_a= temperature of arterial blood (^{0}C), T = tissue temperature (^{0}C).

2.1. Boundary Conditions

Heat loss takes place from the surface of the human body due to convection, radiation, and evaporation because the human body surface is exposed to the environment. So the boundary conditions used in this study is given by

$$R = 0, \qquad \frac{dT}{dr} = 0 \qquad (2)$$

$$R = r, \qquad -K\frac{dT}{dr} = h_C(T - T_\infty) + LE \qquad (3)$$

Where R is the radius of concerned tissue (m), h_C is the coefficient of heat transfer on the surface of the tissue (W/m^2.^{0}C), T_∞ is atmospheric temperature (^{0}C), L is the latent heat (J/Kg) and E denotes the evaporation rate (Kg /m^2.sec).

3. Analytic Solution

Being human body cylinder in shape, equation (1) has been converted into the cylindrical form. The one dimensional steady–state equation in the radial direction is expressed as,

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dT}{dr}\right) + \frac{M}{K}(T_a - T) + \frac{S}{K} = 0 \qquad (4)$$

To perform the non-dimensionalization of (4) with boundary conditions (2) and (3), we introduce the following characteristic quantities and dimensionless parameters,

$$\tilde{r} = \frac{r}{R}, \text{ and } \tilde{T} = \frac{T - T_\infty}{T_a - T_\infty}, \qquad (5)$$

$$\widetilde{M} = \frac{MR^2}{K}, \tilde{S} = \frac{SR^2}{K(T_a - T_\infty)}, \qquad (6a)$$

$$\widetilde{h_C} = \frac{h_C R}{K}, \qquad \widetilde{N} = \frac{LER}{K(T_a - T_\infty)} \qquad (6b)$$

Differentiating **(5)** with respect to 'r' then, substituting in (4) we get

$$\frac{1}{\tilde{r}}\frac{d}{d\tilde{r}}\left(\tilde{r}\frac{d\tilde{T}}{d\tilde{r}}\right) + \frac{R^2 M(1-\tilde{T})}{K} + \frac{SR^2}{K(T_a - T_\infty)} = 0 \qquad (7)$$

With the use of **(6a)** and **(6b), (7)** reduces after calculating to

$$\frac{1}{\tilde{r}}\frac{d}{d\tilde{r}}\left(\tilde{r}\frac{d\tilde{T}}{d\tilde{r}}\right) - \widetilde{M}\tilde{T} + \left(\widetilde{M} + \widetilde{N}\right) = 0 \qquad (8)$$

For computational simplicity, again we put

$$\boldsymbol{M} + \boldsymbol{N} = \text{U}$$

$$\widetilde{\boldsymbol{M}} = \text{V}$$

$$\emptyset = \text{U} - \text{VT}$$

And get the **(9)** as follows

$$\tilde{r}^2 \frac{d^2\emptyset}{d\tilde{r}^2} + \tilde{r} \frac{d\emptyset}{d\tilde{r}} - V\tilde{r}^2\emptyset = 0 \quad (9)$$

This equation is the Modified Bessel's equation of zero order and comparing it with Modified Bessel equation

$$\left[\begin{array}{c} x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} - (\beta^2 x^2 + p^2)y = 0 \\ Whose\ solution\ is\ given\ by \\ y = AI_p(\beta x) + BK_p(\beta x) \end{array}\right]$$

The solution of equation (9) for $\emptyset$ can be written as

$$\emptyset = c_1 I_0\left(\sqrt{V}\tilde{r}\right) + c_2 K_0(\sqrt{V}\tilde{r})$$

Where c_1 and c_2 are arbitrary constants.

After differentiating $\emptyset$ with respect to $\tilde{r}$, calculating and substituting the corresponding values with boundary conditions, the solution for $\tilde{T}$ and then the solution for T can be written as in (10) and (11).

$$\tilde{T} = \frac{U}{V} - \frac{1}{V}\left[\left\{\frac{\widetilde{h_c}U + \tilde{N}V)I_0(\sqrt{V}\tilde{r})}{\widetilde{h_C}I_0(\sqrt{V}\tilde{r}) + \sqrt{V}I_1(\sqrt{V}\tilde{r})}\right\}\right] \quad (10)$$

$$T = T_\infty + (T_a - T_\infty)\left[\frac{U}{V} - \frac{1}{V}\left\{\frac{(\widetilde{h_C}U + V\tilde{N})I_0((\sqrt{V}\tilde{r}))}{\widetilde{h_C}I_0(\sqrt{V}\tilde{r}) + \sqrt{V}I_1(\sqrt{V}\tilde{r})}\right\}\right] \quad (11)$$

Table 1: Thermo-physical parameters [1, 14]

Parameters	Symbols	Values	Unit
Thermal conductivity	K	0.48	W/m^{0}C
Blood specific heat	c_b	3850	J/Kg^{0}C
Blood density	ρ_b	1000	Kg/m^3
Perfusion rate	w_b	3	Kg/s.m^3
Metabolism	S	1085	W/m^3
Arterial temperature	T_a	36.98	^{0}C
Tissue thickness	R	0.03	M
Heat transfer coefficient	h_c	10.023	W/m^2.^{0}C
Latent heat	L	2400000	J/Kg
Evaporation rate	E	0.00004	Kg/m^2.sec
Environmental temperature	T_∞	30	^{0}C

The effect at various temperature profiles based on the above parameters values are presented graphically and discussed below.

4. Results and Discussions

The analytical solution of the bioheat equation for the cylindrical body and its respective results of physical properties depends on various factors. So the following values of parameters from table 1 have been used in equation **(**11) to observe the effect of different atmospheric temperature, thermal conductivities, blood perfusion rates, metabolic heat generation, and heat transfer coefficients.

4.1. Effect of Atmospheric Temperature

The temperature profiles at a various atmospheric temperature below 37^{0}C (normal body temperature) at $T_\infty = 25^0C$, 28^0C, 31^0C and 34^0C, and above 37^{0}C at $T_\infty = 38^0C$, 41^0C, 44^0C and 47^0C, are considered and shown respectively in Figures 1(a), 1(b) and 1(c), 1(d).

Although the graphs in both Figures 1(a) and 1(b) show that the body temperature decreases from the core of the body towards the skin surface if the atmospheric temperature is less than the body core temperature, the graphs in Figure 1(b) show that the body temperature decreases more quickly towards the skin surface than in Figure 1(a). This is due to the evaporation effect. On the other hand, if the atmospheric temperature is higher than the body core temperature, the body temperature increases from the core of the body towards the skin surface, which can be seen in Figures 1(c) and 1(d). Moreover, the graphs in Figure 1(d) show that the temperature increases more slowly towards the skin surface than in the graphs in Figure 1(c) due to sweat evaporation, which helps to cool down the body temperature.

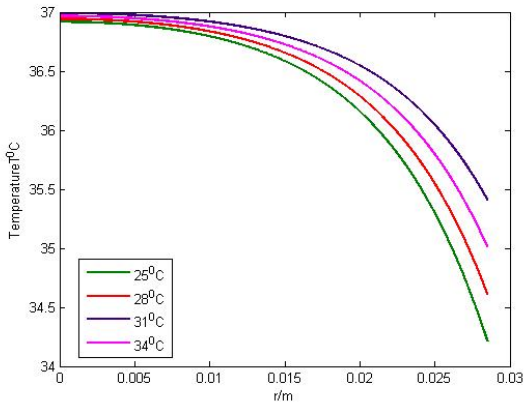


Figure 1(a): Effect of T_∞ < 37^{0}C when E= 0

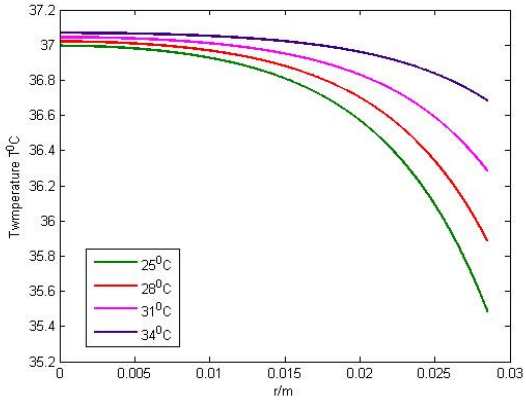


Figure 1(b): Effect of T_∞ < 37^{0}C with *LE*

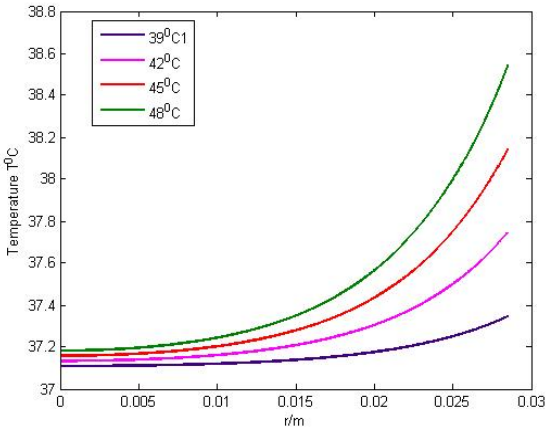


Figure 1(c): Effect of $T_\infty > 37^0$C when E= 0

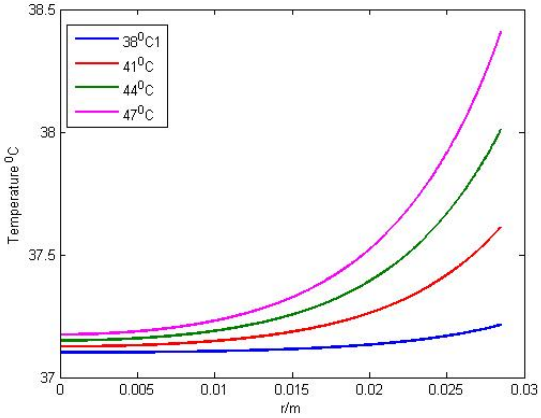


Figure 1(d): Effect of $T_\infty > 37^0$C with *LE*

4.2. Effects of Thermal Conductivities

The temperature profiles at various thermal conductivities K = 0.24W/m^{0}C, 0.48W/m^{0}C, 0.60W/m^{0}C and 0.72W/m^{0}C are shown graphically in Figures 2(a) and 2(b). With the increase in thermal conductivities, the inner part of body temperature increases but the temperature on the body surface decreases sharply as the thermal conductivities increase. It happens due to the conduction process at the body surface.

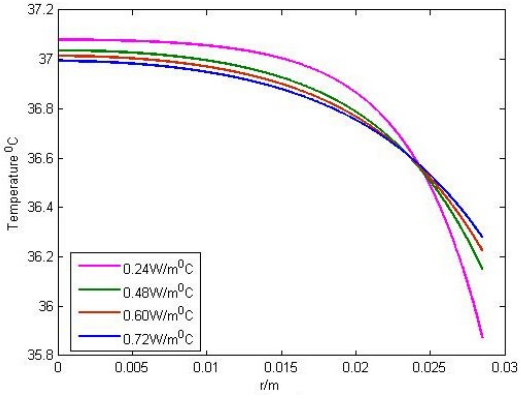


Figure 2(a): Effect of thermal Conductivities when E= 0

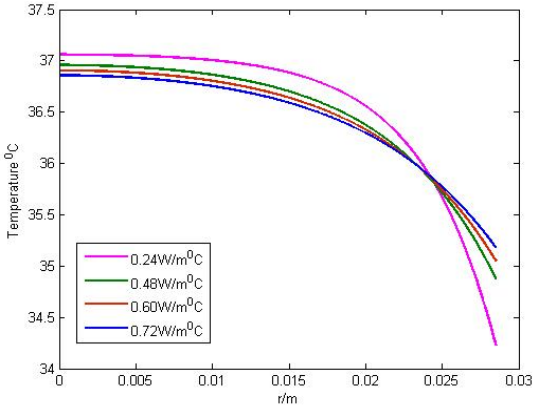


Figure 2(b): Effect of thermal Conductivities with *LE*

4.3. Effect of Blood Perfusion

The various values of blood perfusion are taken as w_b = 1.5 Kg/s.m^3, 2 Kg/s.m^3, 2.5 Kg/s.m^3 and 3 Kg/s.m^3 to observe their effect on the temperature of the body. The graphs in Figures 3(a) and 3(b) indicate that the gradient temperature variation in radial direction decreases with the increase of the blood perfusion. Figure 3(b) shows that the temperature falls more sharply towards the surface of the body than in Figure 3(a) because of the evaporation effect together with the blood perfusion.

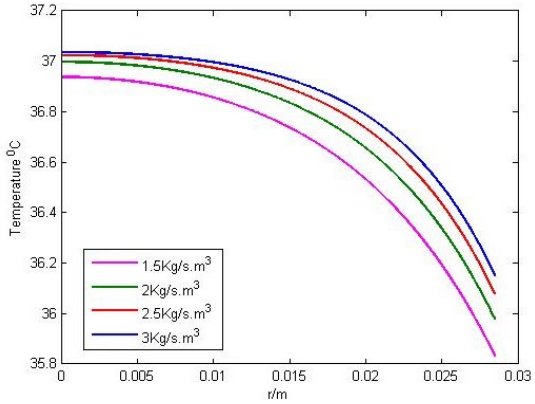


Figure 3(a): Effect of blood perfusion when *E= 0*

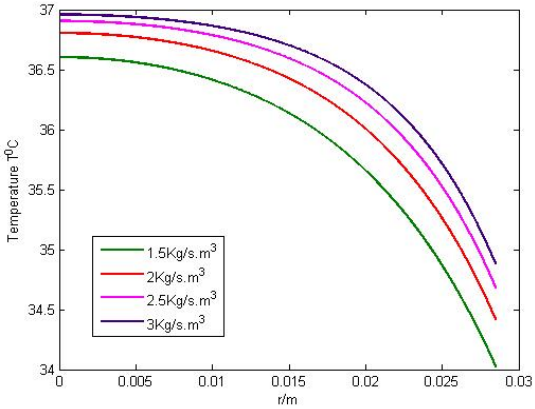


Figure 3(b): Effect of blood perfusion with *LE*

4.4. Effect of Metabolic Heat Generation

The various values of metabolic heat generation q_m = 141.05 W/m^3, 271.25 W/m^3, 542W/m^3 and 1085 W/m^3 are taken. Graphs in Figures 4(a) and 4(b) show that the values of metabolic heat generation have a very small effect on temperature distribution in the human body. The change in metabolic heat from 542W/m^3 to 1085 W/m^3 make the small change in body temperature almost 0.1^{0}C whereas its effect on the skin surface is negligible. This happens due to the negligible concentration of blood vessels towards the skin surface. In the case when sweat evaporation is present (Figure 4(b)), the effect of metabolic heat is almost negligible.

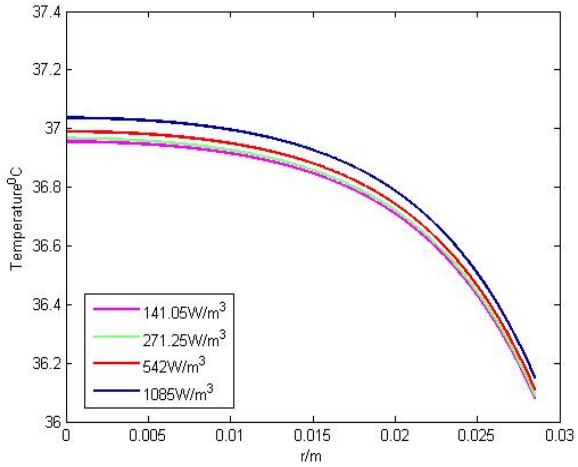


Figure 4(a): Effect of metabolism when $E = 0$

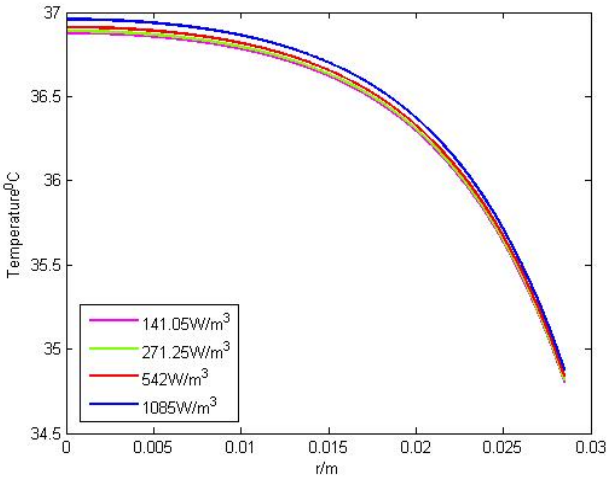


Figure 4(b): Effect of metabolism with *LE*

4.5. Effect of Heat Transfer Coefficients

The various values of heat transfer coefficients are considered at h_C = 5.023 W/m^2.^{0}C, 10.023 W/m^2.^{0}C, 15.023 W/m^2.^{0}C and 20.023 W/m^2.^{0}C. The significant effect of the heat transfer coefficient can be seen in Figures 5(a) and 5(b). The curves in these figures indicate that the gradient temperature in the radial direction decreases with the increase of the heat transfer coefficients.

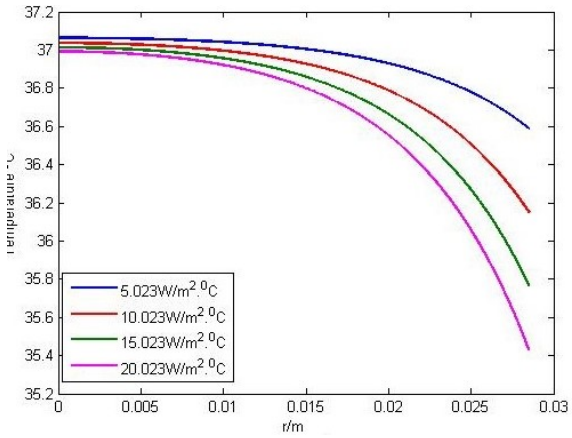


Figure 5(a): Effect of h_C when $E = 0$

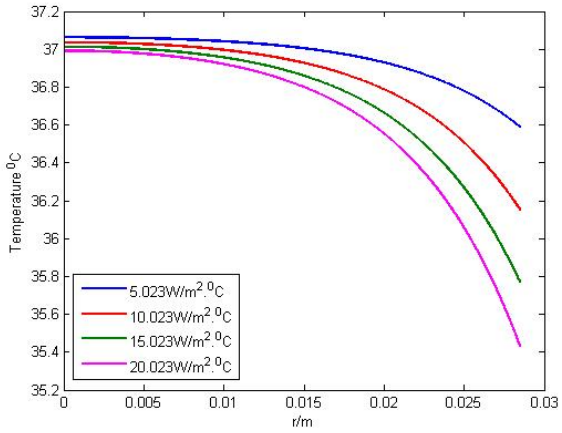


Figure 5(b): Effect of h_C with *LE*

5. Conclusion

In the present study, the analytical solution of the cylindrical form of Pennes' bioheat equation with boundary conditions including the latent heat of sweat evaporation in the one-dimensional steady-state case is obtained. The effect of the temperature changes in various atmospheric temperature, thermal conductivities, metabolic heat generation rates, blood perfusion rates, and heat exchange coefficients has been observed by using the solution obtained from this model. The study reveals that the atmospheric temperature, blood perfusions, and the heat transfer coefficients have a more significant effect on the temperature variation on the skin surface than in the body core. The gradient temperature in the radial direction decreases with the increase of the ambient temperature, the blood perfusion rates, and the heat transfer coefficients. There is a slight variation in body core temperature even an increase in metabolic heat generation. Its effect on the skin surface is negligible. The various parameters used in this model have certainly a more remarkable effect on the temperature distribution in the human body than the case suggested in [1, 2]. Such a model may be useful for the researcher as well as thermal diagnosis and hyperthermia treatment of cancer. The analytic solution obtained in this paper can also be extended in the axial and angular direction as well as unsteady state case.

Conflict of Interest

The authors make it sure that there is no conflict of interest.

References

[1] Kabita Luitel, "Mathematical Model for Temperature Distribution in Cylindrical Human Body" IEEE 2017, 2nd International Conference on Man and Machine Interfacing (MAMI), IEEE 2017.

[2] Yue Kai Zhang, Fan You Xinxin, An Analytic Solution of One – dimensional Steady – state Pennes Bio-heat.Transfer Equation in Cylindrical Coordinates, Journal of Thermal Science, 13(3),52 - 54, 2004.

[3] Ahmed Lakhssassi, Kengne, and Emmanuel Semmaoui, Modified Pennes' equation modeling bio-heat transfer in living tissues: analytical and numerical analysis, Natural Sciences, B, 1375 - 1385, 2010.

[4] Hoppe PR., Heat Balancing modeling, Institute, and Outpatient Clinic for Occupational Medicine, Experientia. 49:741-746, 1993.

[5] Saraswati Acharya, D.B. Gurung, V. P. Saxena, Mathematical Modeling of Sex-Related Differences in the Sensitivity of the Sweating Heat Responses to Change in Body Temperature, British Journal of Mathematics & Computer Science, 12(4), 1 -11, 2016. www.sciencedomain.or.com

[6] Monteith JL., Latent Heat of Vaporization in Thermal Physiology, Nature New Biology 236: 1996, 1972.

[7] George Havenith, Peter Bröde, Emiel den Hartog, Kalev Kuklane, Ingvar Holmer, Rene M. Rossi, Mark Richards, Brian Farnworth, and Xiaoxin Wang, Evaporative cooling: the effective latent heat of evaporation in relation to evaporation distance from the skin, J Appl Physiol, 114, 778 - 785, 2013. doi:10.1152/japplphysiol.01271.2012.

[8] H Arkin., Lx Xu, and K.R Holms, Recent Developments in Modeling, Heat transfer in blood Perfused Tissue, IEEE Transactions on Biomedical Engineering, 41, 97 - 107, 1994.

[9] H.H. Pennes', Analysis of Tissue and Arterial Blood Temperatures in Resting human forearm, Journal of Applied Physiol, 1, 93 - 1948.

[10] M.A. Khandey and V. P. Saxena, FEM Estimation of One-dimensional Unsteady-State Heat Regulation in Human Head exposed to the Cold Environment, Journal of Biological system, 17, 853 -- 863, 2009.

[11] D.B Gurung, V.P Saxena, and P.R Adhikari, Finite Element approach to The one-dimensional Steady-state Temperature Distribution in the Dermal parts with Quadratic Shape Function, J.Appl. Math and information, 27, 301 - 313, 2009.

[12] Mir Aijar and Javid Gani Dar, Mathematical Analysis of Bioheat Equation for the Study of Thermal Stress on Human Brain, Applied Mathematics & Information Sciences and Letters An International Journal, l 5(1), 33 - 39, 2017.

[13] ER Nadel, RW Bullard, JAJ Stolwijk. Importance of skin temperature in the regulation of sweating. J. Appl. Physiol., 46, 430 - 437, 1979.

[14] Lalif M. Jiji, Heat Conduction, 3rd Edition Springer, 52 – 54, 2009.

[15] W.J Minkowycz and E.M, Sparrow Advance in Numerical Heat Transfer Printed in the United States of America, (III), 2009.

[16] Kabita Luitel, D.B. Gurung, Mathematical Model for temperature Distribution in Cylindrical Human Body, The Nepali Mathematical Sciences Report, 3, (1&2), 19 - 28, 2012.

[17] Mir Aijar and M. A. Khanday, Temperature Distribution and Thermal Damage of Peripheral Tissue in Human Limbs During Heat Stress, A Mathematical Model, Journal of Mechanics in Medicine and Biology, 16(2), 16500--16517, 2016.

[18] Ying He, A Numerical coupling model to analyze the blood flow temperature and oxygen transport in Human Breast Tumor and laser irradiation, *Japanese Journal of Computers in Biology and Medicine,* (2006), 36, P.1336-1350.

[19] D.B Gurung, Saraswoti Acharya, Five Layered Temperature distribution in Human dermal Part. Nepali Mathematical Sciences Report, Central Department of Mathematics, Tribhuvan University, Kirtipur, Nepal, 2011.

[20] D.B Gurung, Kabita Luitel, Development of Bio-Heat Equation and its Application, Proceedings of Nepal Mathematical Society, 85 - 94, 2012.

[21] Mamata Agrawal, Neeru Adlakha, R.K Pardadsani, Modeling and Simulation of the thermal effect of Metastasis Timors in Human Limbs, International Symposium on Devices MEMS, Intelligent system and communication (ISDMISC), 24 - 29, 2011.

[22] Gokul K C, D.B. Gurung, and P. R. Adhikari, Thermal Effect of the eyelid in human eye temperature model, J. App. Math. & informatics 32(5-6), 649 - 663, 2014.

[23] Luisiana X. Cundin, William P. Roach, Nancy Millenbaugh, Empirical Comparison of Pennes' Bio-Heat Equation, Proc. Of SPINE, 7175, 717516-1 -- 717516-9.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 438-445 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Study on CD ROADM Contention Blocking

Guangzhi Li[*,1], Kerong Yan[2], Li Huang[2], Bin Xia[2], Fanhua Kong[2], Yang Li[2]

[1]*Futurewei Technologies, Inc., 400 crossing road, Bridgewater, NJ 08807, USA*

[2]*Huawei Technologies Co., Ltd., Shenzhen, 518129, China*

A R T I C L E I N F O

Article history:
Received: 18 October, 2018
Accepted: 05 December, 2018
Online: 19 December, 2018

Keywords:
CD ROADM
Contention Blocking
Contention Avoidance

A B S T R A C T

Service providers are transferring their static optical transport networks from semi-permanent connections to agile automatic switched optical networks (ASON) with dynamic optical connection provisioning and restoration. To achieve this goal, service providers are looking for flexible optical network ROADMs with CDC capabilities. Although many contention scenarios during network connection provisioning and restoration have been illustrated, surprisingly academic simulations have showed that the blocking probability improvement of CDC ROADM comparing with CD ROADM is not significant. This is good news for service providers since most deployed optical networks are only CD ROADM capable, instead of CDC ROADM capable. How to make use of existing CD ROADMs to achieve network automation becomes an urgent challenge. In this paper, we present two research results to attack this challenge: (1) first, we built an analytical mode to estimate the CD ROADM contention blocking probability and show that when a CD ROADM add/drop local direction capacity occupation ratio is low or moderate, the contention blocking probability is not significant. From this model, we estimate that one can use a CD add/drop local direction capacity occupation ratio up to 75% before installing another CD ROADM add/drop local direction or installing a CDC add/drop local direction when available. Simulation results on real network topologies and traffic matrices verified our recommendation; (2) second, we observed that most deployed optical networks are usually providing 100G or 200G per wavelength while majority applications are still requesting much smaller bandwidths and service providers often provide OTN (Optical Transport Network) over ROADM architecture for transport services. Since OTN provides electronic switching capability, in this paper, we present a new algorithm and methodology to make use of both OTN switch and CD ROADM to avoid service contention without using CDC ROADM.

1. Introduction

Color-less, Direction-less, and Contention-less Reconfigurable Add/Drop Multiplex (CDC-ROADM) architectures have recently generated considerable interests among service providers and optical transport vendors [1~7]. There are many technical papers and industry white papers to describe the benefits and applications of CDC ROADM. Comparing with CD ROADMs (Colorless and Directionless only), CDC ROADMs are able to offer additional flexibility and simplicity for optical wavelength planning and operation, especially for dynamic wavelength traffic as well as wavelength dynamic restoration. However CDC ROAMD brings extra components and complexity, which increases ROADM cost. In field network deployments, the majority deployed optical networks are still CD ROADM only. Then following questions come to our mind when service providers upgrade their optical ROADM networks: (1) where does the CD ROADM contention come from? (2) How much is CD ROADM contention blocking? (3) Can we use an analytical formula to estimate CD ROADM contention blocking? (4) Most importantly, what should we do when CD ROADM contention blocking occurs?

In this paper, we present our two research results to answer above questions: (1) first, we built an analytical mode to estimate the CD ROADM contention blocking probability and show that when a CD add/drop local direction capacity occupation ratio is low or moderate, the contention blocking probability is not

[*]Guangzhi Li, Futurewei Technologies, Inc., 400 crossing road, Bridgewater, NJ 08807, USA, Guangzhi.li@huawei.com

significant [8]. From this model, we estimate that one can use a CD add/drop local direction capacity occupation ratio up to 75% before installing another CD add/drop local direction or installing a CDC add/drop local direction when available. Simulation results on real network topologies and traffic matrices verified our findings; (2) second, we observed that currently most optical networks are usually providing 100G or 200G per wavelength while majority applications are still requesting much smaller bandwidths and service providers often design OTN (Optical Transport Network) over ROADM architecture for transport services. Since OTN provides electronic switching capability, in this paper, we present a new algorithm and methodology to make use of both OTN switch and CD ROADM to avoid service contention without using CDC ROADM.

2. ROADM Architecture Comparison

Early stage of optical transport network is typical point-to-point wavelength division multiplex (WDM) system which consists of two terminals connected by a pair of fibers. Each terminal contains an optical wavelength multiplexer and de-multiplexer, amplifier, and transponders that interface client signals. The number of inline amplifiers placed between the terminals depends on the length and quality of fiber. Each transponder re-transmits its incoming client signal onto a particular wavelength of the optical grid (also called a channel) and the optical multiplexer combines these signals at different wavelengths together and transmits the combined signal over a fiber to the de-multiplexer at the other end. The de-multiplexer decomposes the multiplexed signal back into original signals at their respective wavelengths, which are re-transmitted by each receiving transponder into its client signal.

When a point-to-point demand (also called connection) is transported over two WDM systems, two OTs are needed: the first OT converts the first WDM system wavelength into the common short-reach wavelength (λ0), and the second OT converts the λ0 short reach wavelength into the second WDM system wavelength. However, when the two WDM systems are of the same technology from the same vendor, one regenerator can be used to replace the two back-to-back OTs, avoiding the conversion to the common short-reach wavelength, and thus reducing the component costs.

Later on, Reconfigurable Optical Add Drop Multiplexer (ROADM) technologies for optical transport network have been deployed due to their high capacity and capital savings. A ROADM network typically includes a set of multi-degree nodes connected via fibers to form a mesh topology. Traffic may be added or dropped, regenerated, or expressed through at ROADM nodes. Inside ROADM node, each network side WSS (wavelength selective switch) is called one line direction, each add/drop side WSS is called add/drop local direction. At the add/drop local direction, the classic ROADM was designed with fixed wavelength transponders and directed tributary for each line direction, and it was called colored and directed ROADM. Each transponder only is allowed to transmit signal on a fixed wavelength and to a fixed direction, see Figure 1 for an example. This ROADM design is acceptable for static optical connections. Network planners could plan the connections carefully and deploy them as planned for quite a long time. As network traffic grows and become more and more dynamic, service providers prefer to automate optical connection provisioning without manual intervention. In such an operation scenarios, this colored and directed ROADM architecture is no longer able to satisfy this requirement. For example, to reroute an existing wavelength to a different direction or to reuse an existing transponder for a different wavelength, all require new ROADM add/drop architecture, which was colorless and directionless ROADM, the so called CD ROADM.

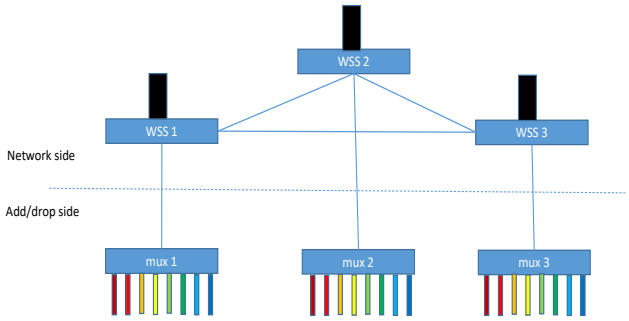


Figure 1: Example of colored and directed ROADM

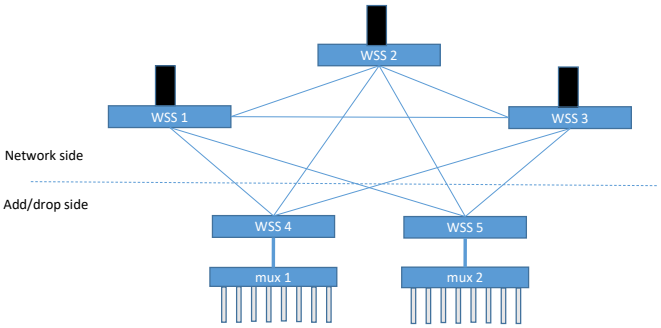


Figure 2: Example of CD ROADM

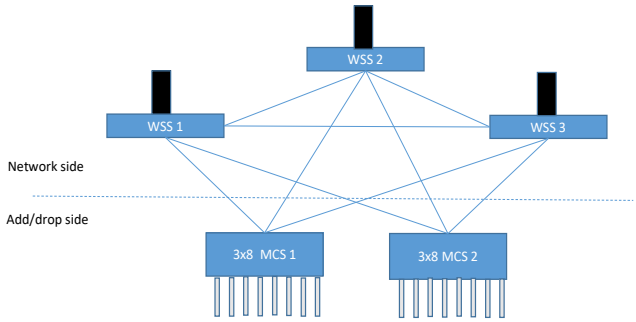


Figure 3: Example of CDC ROADM

Figure 2 shows an example of CD ROADM architecture with 3 line directions and 2 add/drop local directions. There are 8 tunable transponders, which means each transponder can be tuned to any wavelength. However, no more two transponders at the same add/drop local direction can be tuned to the same wavelength since all tuned wavelengths from transponders will be multiplexed into a single fiber pair. This limitation is called add/drop local direction contention. If two clients are connected to two transponders at a common add/drop local direction. Each client has a wavelength request at the same time, and routing wavelength assignment algorithm finds the same wavelength for the two requests. Due to add/drop local direction contention, the

two transponders could not use the same wavelength. Then either one of them has to find a different wavelength or may be blocked. This limitation of CD ROADM architecture leads to another advanced colorless directionless, contention-less ROADM architecture, the so called CDC ROADM.

Figure 3 shows an example of CDC ROADM architecture with 3 line directions and 2 local directions. In this architecture, the WSS (wavelength selective switch) and Multiplex are replaced by a 3x8 multicast optical switch (MCS). Without the fiber bottleneck of CD architecture, the transponders in same add/drop local direction can be tuned to the same wavelength, i.e., there is no wavelength contention in CDC ROADM architecture. This wavelength assignment flexibility is brought in by the relatively expensive component of MCS. In Figure 3, one can tune 3 transponders at a single add/drop local direction into a single wavelength since there are total 3 line directions. Of course, it is unnecessary to tune more than 3 transponders into a single wavelength since there are at most 3 common wavelengths can be supported at the network side. This extra flexibility should be able to improve network capacity utilization and reduce network blocking probability.

3. CD ROAMD contention

It is widely accepted that CD ROADM architecture causes wavelength contention during wavelength provisioning when two connections with the same wavelength need to be added/dropped at the same add/drop local direction [9~12]; during wavelength restoration when the network available wavelength has been used by other connection at the same add/drop local direction, and/or during regeneration when the connection wavelength is free at most one add/drop local direction only. To overcome CD ROADM contention issue, one may add as many local directions as the number of ROADM line directions. Then if there is one free wavelength at any line direction, there must exist one add/drop local direction with the same wavelength free. However during wavelength restoration, contention could still occur: during wavelength restoration, the transponder at the client side could be reused, and the restoration wavelength at network side is available while the restoration wavelength may not be available at add/drop local direction of the client transponder. To solve this contention, client-side optical cross connect architecture is recommended [10]. Problem seems solved, the issue is the extra cost of cross-connect. At each ROADM site, the number of connection requests usually much less than the value of M*W, where M is the number of line directions and W is the number of wavelengths per fiber. Large number of add/drop local directions will reduce the available number of potential line directions for the same size of WSS, which makes network expansion difficult. In real deployed optical networks, the number of add/drop local directions usually is much smaller than the number of line directions. In this case, an expensive CDC add/drop local direction could be used to avoid wavelength contention.

All above analysis is quantitative with extreme case assumptions. How severe is the CD ROADM contention during optical network planning and operation? Academic studies show that under reasonable assumption on network topologies and demands, the blocking probability improvement using CDC ROADM comparing CD ROADM with the same number of add/drop local directions at each node is not significant [13~15]. This surprising observations brought us interests for a further investigation at CD ROADM contention issue, including theoretical analysis and simulation verification, as well as how to resolve CD ROADM contention.

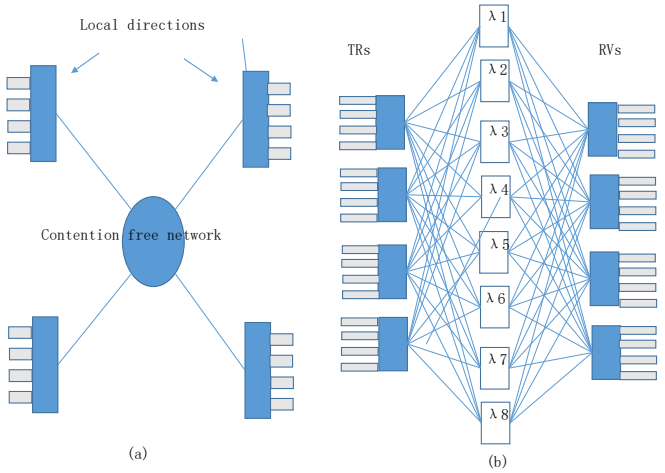


Figure 4: Clos model of CD ROADM

A connection in a CD ROADM network is blocked when the network could not find a free wavelength, or the wavelength is not available at either side of add/drop local directions. The first one is due to optical network wavelength continuity constraint and the second is due to CD wavelength contention. Without considering network wavelength continuity constraint, we could simply model the CD ROADM network as a star topology, see figure 4(a): the network is a full contention free optical switch at the center, each CD add/drop local direction is a star terminal. Assume there are n tunable transponders at each terminal, and each star branch fiber supports m wavelengths. Each transponder can be tuned to any wavelength, but no more two transponders can be tuned to a same wavelength. Then two terminals are able to establish a connection if and only if the two terminals have two transponders and the two branch fibers have a common free wavelength. If there are total r add/drop local directions, the simplified star model can be viewed as a (r,n,m) clos network [16], see figure 4(b), where a single transponder is separated into one transmitter (TR) and one receiver (RV). Thus the non-blocking condition would be $m \geq 2n-1$, i.e., $n \leq (m+1)/2$. So when an add/drop local direction is half filled, in order to avoid contention, it would be the time to add new CD add/drop local directions, or upgrade CDC add/drop local directions. Of course this simplified model has drawbacks since the optical network is not contention free at all. Even if two add/drop local directions have free transponders and same available wavelengths, but the network may not have the same wavelength between the associated two nodes; on the other hand, if the associated two node has an available wavelength along a path between them, the two add/drop local directions may not have the same common wavelength available. Thus our model is under estimate the contention blocking. However since there could be large number of paths between any two network nodes in real networks, we think this estimation may not be too off from reality. In next sections, we will verify our simplified model via simulation comparing CD ROADM and CDC ROADM of real networks with real demand matrix.

Academic papers show that even if the CD add/drop local direction traffic is more than half-filled, the blocking probability

is still relatively low. When we want to create a connection between two CD add/drop local directions, again we look at the clos network model, which means we want to establish a connection from one TR bank *I* of the first stage to one RV bank *J* of the third stage. We define clos network state *{u,v}* of TR bank *I* and RV bank *J* as *u* TRs are busy and *v* RVs are also busy, and define *B(u,v)* as the contention blocking probability in state *{u,v}* independent of other TR banks and RV banks.

According to Jacobaeus [17], the contention blocking probability *B(u,v)* in state *{u,v}* is:

$$B(u,v) = \begin{cases} 0 & when\ u+v < m \\ \dfrac{u!\,v!}{m!\,(2n-m)!} & when\ u+v \geq m \end{cases}$$

Simple proof: when $u+v < m$, there is at least one wavelength available in the second stage, which is reachable for both TR bank *I* and RV bank *J*, so no blocking for a new connection. When $u+v \geq m$, depending on how *u* and *v* connect to the second stage wavelengths, there could be no more wavelength available for next connection if and only if RV bank *J* all available wavelengths *(m-v)* resides in TR bank *I* busy wavelengths. So the blocking probability could calculated as:

$$B(u,v) = \frac{choosing\ m-v\ wavelengths\ from\ u\ wavelengths}{choosing\ m-v\ wavelengths\ from\ m\ wavelengths} = \binom{m-v}{u} \Big/ \binom{m-v}{m} = \frac{u!v!}{m!(u+v-m)}.$$

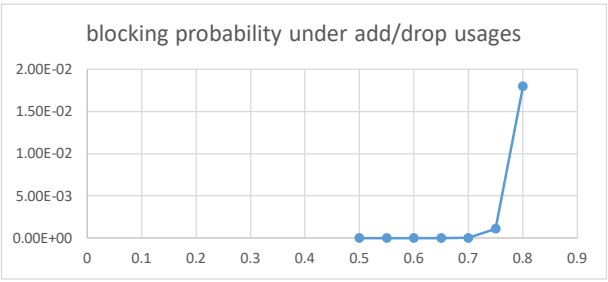


Figure 5: CD ROADM estimated blocking probability

Now we let $u=v=x*m$, where $m = 80$, and calculate the blocking probability of $x=0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8$. See Figure 5, where x-axis is the value of *x* for number of existing connections in CD add/drop local directions, and y-axis is the estimated blocking probability. It is easy to see that even if the add/drop local direction is 80 percent full, the contention blocking probability is still under 2% without considering network wavelength continuity constraint. When the CD add/drop dimension is 75% full or less, the contention blocking probability is still negligible. Based on this analytical modeling, it explains why all academic simulations reported very small blocking probability using CD ROADM architecture. In next sections, we will use real network topology and network traffic matrix to simulate CD ROADM blocking probability. Our results confirm this analysis with very small variance.

4. An example interpretation of clos model

In section 3, we model CD ROADM as a clos network model and claim that when the CD ROADM local directions are not half filled, the CD ROADM blocking probability will be similar to CDC ROADM blocking probability. This claim may not be easy to interpret. In this section, we give one simple example to show that when CD ROADM local direction is half-filled, some CD ROADM connections will be blocked while the same network configuration with CDC ROADM will not block those connections.

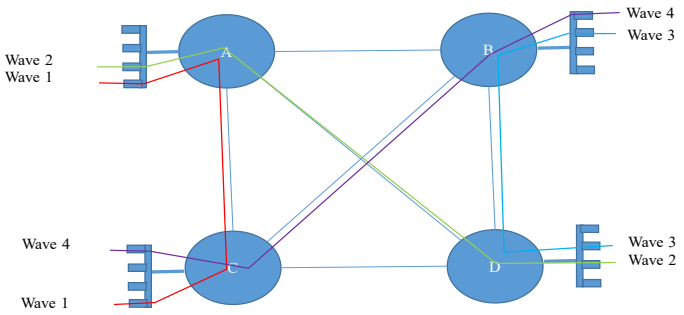


Figure 6: Example of CD ROADM half-filled blocking

Figure 6 shows a 4 node ROADM network and each node is deployed with a single CD add/drop local direction. Without loss of generosity, we assume that each fiber supports 4 wavelengths. Due to network dynamic operation and wavelength connections arrive and leave. At some network stage, we assume there are only 4 connections: connection 1 from node A to node C with wavelength 1; connection 2 from node A to node D with wavelength 2; connection 3 from node B to node D with wavelength 3; and connection 4 from node B to node C with wavelength 4. If there is a new connection request from node A to node B, the request will be blocked due to CD ROADM local contention, while CDC architecture will be able to provision such connection request; similarly another connection request from node C to node D will be blocked also in CD architecture while CDC architecture will not block it. In this simple example, it is easy to see that CD ROADM local direction half-filled is a critical point to cause contention blocking.

5. CD/CDC ROADM network simulation

We first simulate a few planed or planning CD ROADM optical networks under preplanned demand matrixes with and without rerouting restoration. Table 1 shows the parameters of these networks, where #degree is calculated as *2*(#links)/(#nodes)*, where *#links* means the number of network links, and *#nodes* means the number of network nodes. The #add/drop direction distribution *X(Y)* means *y* network nodes with *x* add/drop local directions. We first assume all add/drop local directions are CD architecture, then using a multipath routing and first-fit wavelength assignment algorithm to provision the connections as many as possible.

Table 1: Simulated network parameters

networks	#node	#link	#degree	#connections	#add/drop direction distribution X(Y)
Net 1	30	71	4.73	324	2(2),1(18),0(10)
Net 2	28	43	3.07	404	2(5),1(23)
Net 3	55	81	2.95	723	3(2),2(4),1(47)

Table 2: Simulation results of CD vs CDC

networks	CD (W) blocking	CDC (W) blocking	R(CD)	R(CDC)	Extra CD for working	Extra CD for rerouting	Max CD fill ratio before extra CD	Max CD fill ratio after extra CD
Net 1	0	0	0.43%	0	0	8	86.25%	43.50%
Net 2	7.4%	1.7%	16.5%	16.9%	3	3	85%	56.70%
Net 3	3.87%	1.93%	6%	3.75%	12	12	88.80%	53.00%

For those provisioned connections, we simulate single failure rerouting and calculate the maximal success rerouting ratio:

$$R(CD) = \frac{total\ rerouting\ sucess\ connections\ over\ all\ failures}{total\ failed\ connections\ over\ all\ failures}.$$

Then we assume all add/drop local directions are CDC architecture, then using the same algorithm to provision the connections as many as possible. For all provisioned connections, we also simulate single failure rerouting and calculate the maximal success rerouting ratio, R(CDC).

To investigate the CD contention impact, we try to add extra CD add/drop local directions to critical nodes such that the CD blocked demands equal to CDC blocked demands. The CD add/drop local direction adding policy is based on decreasing ratio of *(#add/drop_connections)/(80*#add/drop_direction)* at each network node. We simulated both with and without failure rerouting cases. Table 2 shows our simulation results, where "Extra CD for working" means total extra CD local directions for working only with the same blocking ratio as CDC ROADM architecture, "Extra CD for rerouting" means total extra CD local directions for working and restoration with the same blocking ratio as CDC architecture, "Max CD fill ratio before extra CD" means maximal CD add/drop local direction fill ratio before adding extra CD add/drop local directions, and "Max CD fill ratio after extra CD" means maximal CD add/drop local direction fill ratio after adding extra CD add/drop local direction for rerouting. We observed that for all tested real networks, the differences between CD and CDC are relatively small. Especially when node CD fill ratio (total #add/drop demands divided by link capacity times #add/drop directions) is less than 50%, CD and CDC outputs are almost the same, which means that CDC does not provide any extra benefits in this case; on the other hand, under the same network topology and demand matrix, in order to achieve the same throughput and rerouting success ratio, CDC did require relatively fewer add/drop local directions, but not significant.

All these tests are based on off-line planning algorithms with static traffic demands. In order to simulate dynamic traffic, we also created a hypothetic network with 50 nodes, 123 links with average node line directions about 4. Link capacity is 80, and each node has one add/drop local direction. In this simulation, we assume paths between any two nodes are all reachable and no regenerator is used. Demands are random generated with poison arrival and expectation holding time. We measure the Erlang value when blocking probability is under 1%. We observed that CDC architecture could improve the throughput of Erlangs about 4% comparing with CD architecture, see Figure 7. Again CDC is able to improve dynamic traffic throughput, but the improvement is not significant.

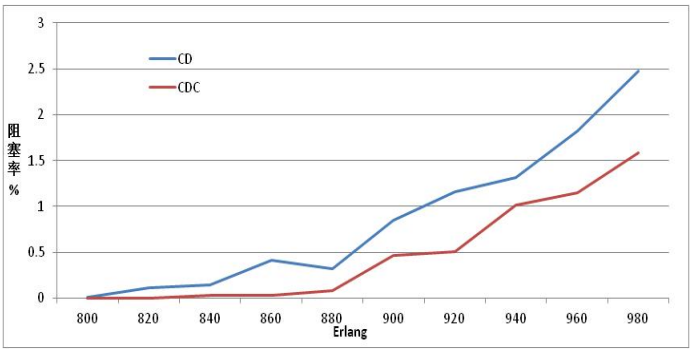


Figure 7: Throughput improvement of CDC vs CD

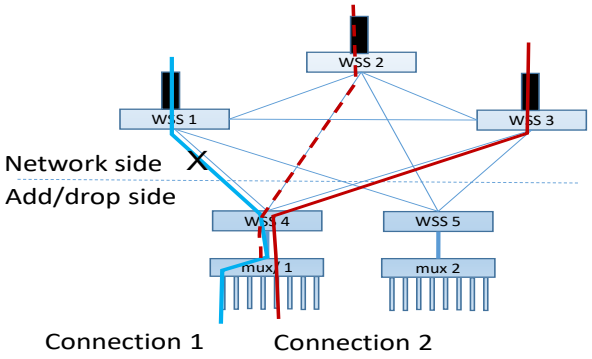


Figure 8: Example of CD ROADM contention

6. A Contention Solution for CD ROADM

Although CD ROADM contention blocking is small when add/drop local direction utilization is low or medium, there is still some possibility of contention blocking during dynamic optical network operation, such as ASON (automatic switch optical network) connection provisioning or restoration. For example in Figure 8, assuming local direction WSS 4 has two connections of 1 and 2, connection 1 was assigned wavelength 1 and routed along WSS 1 line direction while connection 2 was assigned wavelength 2 and routed along WSS 3 line direction. If connection 1 fails and required rerouting restoration, and the source node finds out that

only wavelength 2 along WSS 2 line direction is available and other wavelengths along other WSS line directions are not available. But due to add/drop local direction WSS 4 wavelength contention, wavelength 2 could not be used for connection 1 rerouting, which could lead to connection 1 rerouting failure. If this ROADM is CDC architecture, connection 1 is able to be routed to WSS 2 line direction with free wavelength 2.

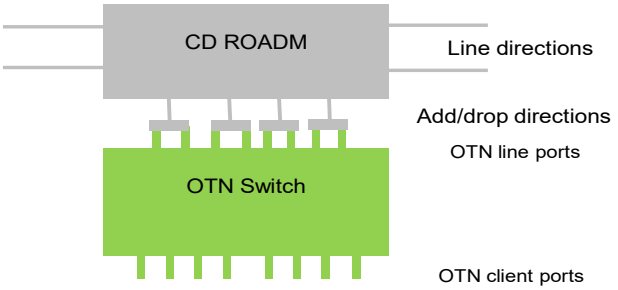


Figure 9: OTN over ROADM architecture

As we mentioned before, CDC ROADM relies on MCS module to multicast each wavelength to each direction, which increases hardware cost and system complexity. In this paper, we present a new solution to avoid CD ROADM wavelength contention during wavelength provisioning and rerouting using existing mature technology without increasing hardware cost. We noticed that most of ROADM node deployments are armed with OTN electronic switching node to form an OTN over ROADM architecture, see Figure 9, where OTN switch is used for traffic grooming and ROADM is used for wavelength switching. However such OTN over CD ROADM architecture still could not be equivalent to CDC ROADM architecture. For example, assume a CD ROADM has 4 line directions, 4 add/drop local directions, and each fiber has 4 wavelengths. Since OTN line ports are pre-installation in field operation, assuming 50% connections add/drop per node, the service provider pre-installed 8 OTN line ports and their usage is based on average policy as recommended by paper [10]. If connection wavelength sequence order is 1,1,1,2,3,3,3,2, the last connection would be blocked due to wavelength contention, however CDC ROADM won't have this contention. However under OTN over ROADM architecture, we have following observations:

Observation 1: wavelength contention blocking condition is that for a given wavelength, all add/drop local directions with free OTN line ports have used the given wavelength.

Observation 2: if CD ROADM has equal number of add/drop local directions as line directions, and one line direction has a free wavelength, then at least one add/drop local direction has the same free wavelength.

Observation 3: given a connection routing wavelength, if the add/drop local direction with the free given wavelength has free OTN line ports connected, the connection can be deployed without contention;

Observation 4: given a connection routing wavelength, assuming equal number of add/drop local directions and line directions, then according to observation 2, at least one add/drop local direction has the wavelength free, say WSS X, and X has no free OTN line port, but add/drop local direction WSS Y has free OTN line port. If OTN line ports are equally connected to CD ROADM add/drop local directions, then WSS X at most has one OTN line port less than WSS Y. So WSS X at least has one wavelength occupied but WSS Y is still free. This is because if WSS Y occupied all wavelengths of WSS X used, and WSS Y also used the given new wavelength, and free OTN line port, then WSS Y would have at least 2 more OTN line ports than WSS X, which contradicts the assumption of OTN line ports equal distribution to ROADM local directions.

Observation 5: if we can change the free OTN line port (say y) of WSS Y into the wavelength that WSS X used but WSS Y is free, say λ1, identify wavelength λ1′ ROADM line direction (say d), cross-connect the OTN line port y to ROADM line direction d, switch the service OTN client port to OTN line port y, free WSS X wavelength λ1 occupied OTN line port (say z), then WSS X has free wavelength λ, and connection free ONT line port z, thus the new connection can be provisioned without contention.

Observation 6: to achieve hitless switching, one could use make-before-break technology; however traditional make-before-break usually is used on the same layer of technology, such as in optical layer or electronic layer, in this proposal, the make-before-break is used in cross-layer between electronic layer and optical layer. We have discussed with both optical system engineers and electronic system engineers and was confirmed that it can be done using existing technologies.

7. Two Use Cases

In this section, we provide two use cases on how to apply the method of section 6 to avoid CD ROADM contention.

7.1. Use case 1: service provisioning contention blocking avoidance

Figure 10 shows an OTN over CD ROADM service provisioning contention blocking avoidance process. Assume the OTN, has 8 client ports numbered from 1 to 8, and 8 line ports connecting to 4 add/drop local directions of a CD ROADM, numbered as x.y, where x is the CD ROADM add/drop local direction number, y is the line port sequential order number at the local direction. For example 2.1 means the first line port at the second local direction of the ROADM.

We further assume that each fiber supports 4 wavelengths, OTN client port 1 cross-connects to line port 1.1 with wavelength 1, client port 2 cross-connects line port 2.1 with wavelength 1, client port 3 cross-connects line port 3.1 with wavelength 2, client port 4 cross-connects to line port 4.1 with wavelength 2, client port 5 cross-connects to line port 1.2 with wavelength 3, client port 6 cross-connects line port 2.2 with wavelength 3, client port 7 cross-connects to line port 3.3 with wavelength 3. At this time, there is still one free client port 8 and one free line port 4.2. If a new service finds available wavelength 2 over one line direction, the service cannot be deployed since CD ROADM local direction 4 already occupied wavelength 2 (see figure 10.a). Contention blocking occurs and following procedure could resolve this contention.

- ✓ Identify free line port: 4.2 on local direction 4
- ✓ Identify one local direction with free wavelength 2: local direction 1.
- ✓ Identify one OTN line port with occupied wavelength from local direction 1 which is free at local direction 4: line port

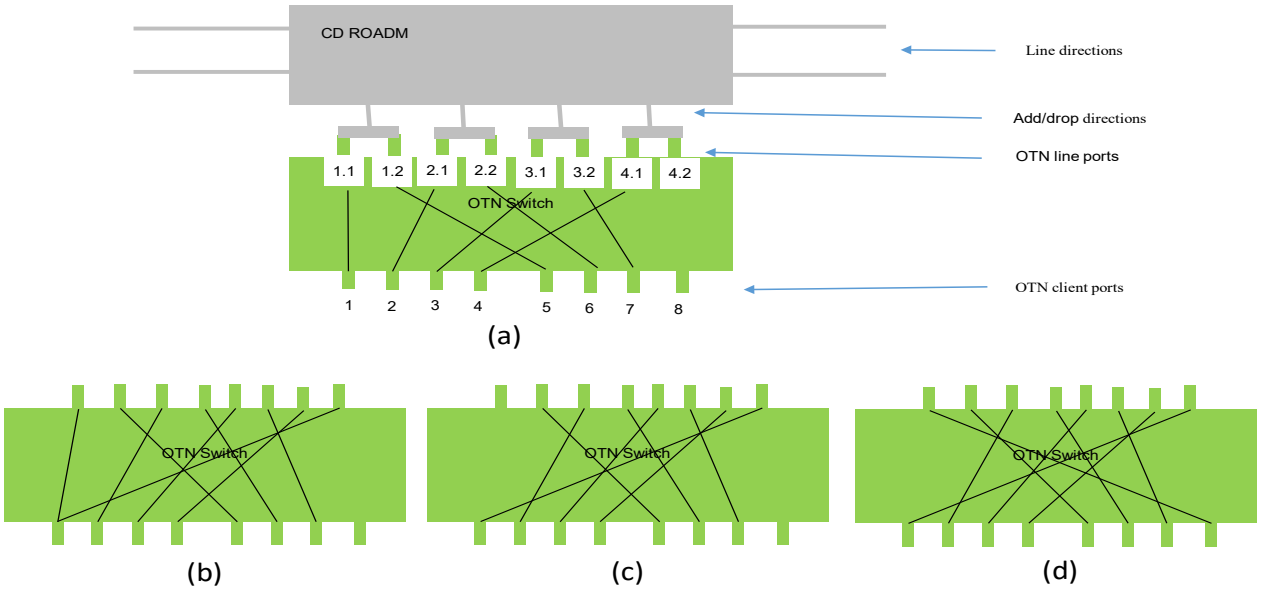


Figure 10: OTN over ROADM provisioning contention avoidance

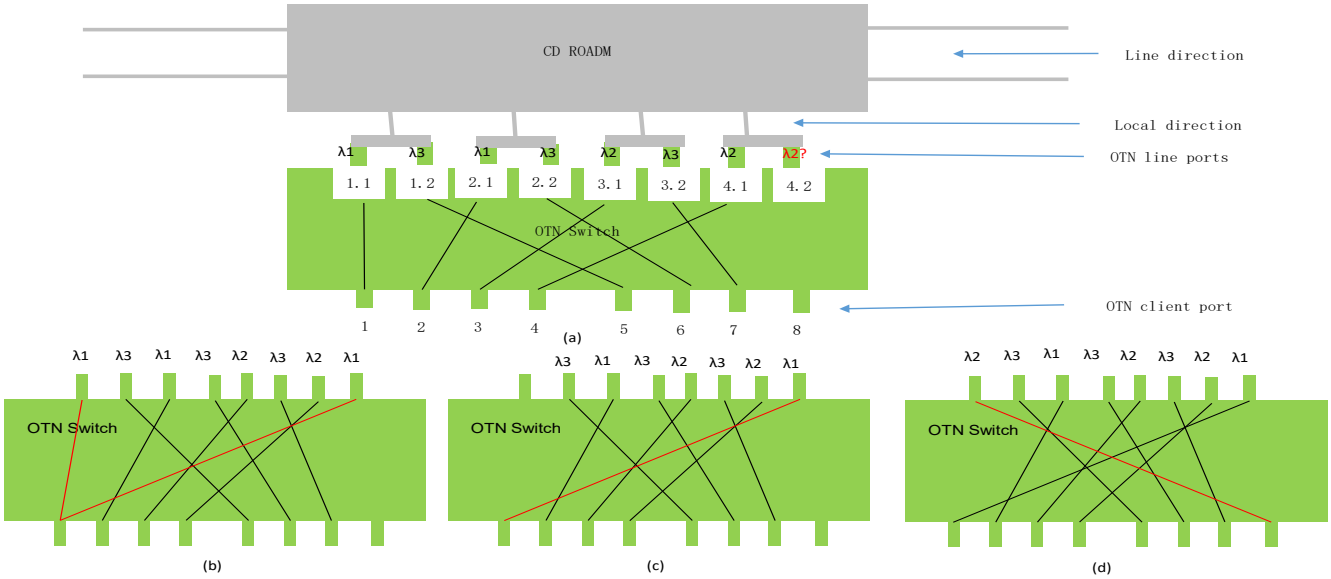


Figure 11: OTN over ROADM rerouting contention avoidance

1.1 with wavelength 1. This can be done according to section 6 observation 4

- ✓ Identify line port 1.1 associated client port: client port 1.
- ✓ Client port dual-cast to line ports 1.1 and 4.2, see figure 10.b
- ✓ CD-ROADM cross-connects 4.2 signal to the same ROADM line direction of line port 1.1
- ✓ ROADM line WSS of line port 1.1 selects signal from line port 4.2
- ✓ Client port cancels signal to line port 1.1, then line port 1.1 is free, see figure 10.c
- ✓ Client port 8 new service cross-connects to line port 1.1 with wavelength 2. See figure 10.d

7.2. Use case 2: failure rerouting contention blocking avoidance

Figure 11 shows the use case of service failure rerouting contention avoidance solution. Figure 11.a shows service OTN internal configuration and wavelength assignment. When a service fails, wavelength rerouting restoration is activated. If rerouting wavelength is the same as original working service wavelength, then OTN configuration has no change, the only change is CD ROADM to cross-connect line port signal from original working line direction to rerouting line direction; If rerouting wavelength is different from original working wavelength, and the local direction of working wavelength has free rerouting wavelength, then the OTN configuration does not need to change, the only change is to tune the working line port to rerouting wavelength and ROADM to cross-connect the rerouting wavelength to rerouting line direction. However if rerouting wavelength is different from working wavelength, and the local direction of working wavelength has occupied the rerouting

wavelength, then we need to find out all local directions with free rerouting wavelength. If one of them has free line port, then we could tune this line port to the rerouting wavelength, ROADM cross-connect the rerouting wavelength to rerouting line direction, and cross-connects the service client port to the rerouting wavelength line port. The difficult case is when all local directions with free rerouting wavelength have no free line ports, we need to adjust at most one existing service configuration. Assume OTN configuration as in figure 11 (a) and client port 7 service fails and rerouting wavelength is wavelength 2. Then working wavelength local direction 3 has no free wavelength 2. After identifying all local directions having free wavelength 2, we find local direction 1 and local direction 2. However both local directions have no free line ports. We choose one local direction with free rerouting wavelength and this local direction at least has one occupied wavelength which is not used by the local direction of failed working wavelength, this can be done according to section 6 observation 4. In Figure 11.a, we choose local direction 2, where wavelength 3 is not used by local direction 3 (failed working wavelength is not used anymore). We reuse OTN line port 3.2, which is free currently and tune it into wavelength 3, duel-cast client port 6 service to both line port 2.2 and line port 3.2, (figure 11.b) cross-connect line port 3.2 signal to ROADM line direction of line port 2.2 connected line direction, and let this line direction WSS selects signal from line port 3.2, cancel signal cast from client port 6 to line port 2.2 and free line port 2.2 (figure 11.c), at last tune line port 2.2 to rerouting wavelength 2 and cross-connects client port 7 signal to line port 2.2, ROADM cross-connects line port 2.2 signal to rerouting line direction (figure 11.d)

8. Conclusion

In this paper, we investigated CD ROADM contention blocking problem. Since ROADM network operation is moving from static to dynamic, from 1+1 failure protection to dynamic rerouting restoration, CDC ROADM architecture has attracted lots of attention. However most deployed ROADM networks are still CD architecture. Is CD ROADM a show stop for dynamic operation? How much contention blocking will CD ROADM produce? If CD ROADM contention happens, is there any solution to solve this contention? We have answered all above questions using analytical model, simulation, and configuration method of procedures. From mathematic model and simulation results, we showed that CD ROADM did produce some contention during network operation, but the contention blocking is not as significant as we originally thought. If the service fill ratio at each local direction is 50% or less, the contention blocking is almost negligible. When contention did occurs during network operation, if service providers are using OTN as grooming and ROADM as wavelength transmission, as the often used architecture in field deployments, we provided a method to reconfigure at most one existing service to avoid the contention blocking. Based on our study, we concluded that a service provider with CD ROADM optical network can move forward to ASON (automatic switched optical network) dynamic operation without worrying too much of CD ROADM contention blocking.

References

[1] Mina Paik, “4 Things we learned at #OFC17”, March 30, 2017, http://www.ciena.com/insights/articles/4-Things-We-Learned-at-OFC17.html

[2] Fujitsu Network Communications Inc.,“CDC ROADM applications and cost comparison”, whitepaper, 2014, https://www.ofcconference.org/getattachment/188d14da-88ba-4a63-91d6-1cc14b335d8b/CDC-ROADM-Applications-and-Cost-Comparison.aspx

[3] Nokia, “Benefits of CDC-F ROADMs”, whitepaper, 2017, https://knect365.com/ngon/article/e081ea00-9bc9-433d-ae21-94617299be41/whitepaper-benefits-of-cdc-f-roadms

[4] Huawei Technologies Co., Ltd., “white paper on technological developments of optical networks”, whitepaper, 2016, http://www-file.huawei.com/-/media/CORPORATE/PDF/white%20paper/White-Paper-on-Technological-Developments-of-Optical-Networks.pdf

[5] Sterling Perrin, “Building a fully flexible optical layer with next-generation ROADMs”, heavy reading whitepaper, 2011, http://www-file.huawei.com/-/media/CORPORATE/PDF/white%20paper/White-Paper-on-Technological-Developments-of-Optical-Networks.pdf

[6] B. C. Collings,, “Advanced ROADM technologies and architectures,” Los Angeles, CA, Paper Tu3D.3, OFC 2015.

[7] S. Poole, S. Frisken, M. Roelens, and C. Cameron, “Bandwidth-flexible ROADMs as network elements,” Los Angeles, CA, Paper OTuE1, OFC 2011.

[8] Guangzhi Li, Kerong Yan, Li Huang, Bin Xia, Fanhua Kong, Yang Li, “How much is CD ROADM contention blocking?”, San Diego, CA, OFC 2018.

[9] L.Zong, H. Zhao, Z. Feng, and S. Chao, “Demonstration of ultra-compact contentionless Roadm based on flexible wavelength router,” ECOC 2014.

[10] M. Feuer, S. Woodward, P. Palacharla, X. Wang, I. Kim, and D. Bihon, “Intra-node contention in dynamic photonic networks”, JOLT, 29(4), Feb. 2011.

[11] Y. Li, L. Gao, G. Shen, and L. Peng, “Impact of ROADM colorless, directionless, and contentionless (CDC) features on optical network performance,” JOCN, 4(11), Nov. 2012.

[12] J. Simmons, “a Closer look at ROADM contention,” IEEE Comm. Mag, 55(2), pp 160-166, Feb 2017.

[13] F. Naruse, Y. Yamada, H. Hasegawa, and K. Sato, “Evaluations of OXC hardware scale and network resource requirements of different optical path add/drop ratio restriction schemes,” J. Opt. Commun. and Netw., 4(11), pp B26-B34, Nov. 2012.

[14] T. Zami, P. Jenneve, and H. Bissessur, “Fair comparison of the contentionless property in OXC,” Asia Commun. and Photonics Conf., Hong Kong, AM3G.3, 2015.

[15] J. Pedro and S. Pato, “Impact of add/drop port utilization flexibility in DWDM networks,” JOCN, 4(11), Nov 2012.

[16] C. Clos, “A study of non-blocking switching networks,” BSTJ, 32(2), pp 406-424, March 1953.

[17] C. Rigault, “Clos networks: a correction of the Jacobaeus result,” Analysis of telecommunications, 57(11-12), pp 1244-1252, Nov 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 446-454 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Real Time Eye Tracking and Detection- A Driving Assistance System

Sherif Said[1,2,*], Samer AlKork[1], Taha Beyrouthy[1] , Murtaza Hassan[1], OE Abdellatif[2], M Fayek Abdraboo[2]

[1]*College of Engineering and Technology, American University of the Middle East, Kuwait*

[2]*Shoubra Faculty of Engineering, Benha University, Shoubra, Cairo, Egypt*

ARTICLE INFO

Article history:
Received: 31 October, 2018
Accepted: 05 December, 2018
Online: 19 December, 2018

Keywords:
Eye Tracking System
Viola-Jones
Haar Classifiers
Smart systems
Driver's safety

ABSTRACT

Distraction, drowsiness, and fatigue are the main factors of car accidents recently. To solve such problems, an Eye-tracking system based on camera is proposed in this paper. The system detects the driver's Distraction or sleepiness and gives an alert to the driver as an assistance system. The camera best position is chosen to be on the dashboard without distracting the driver. The system will detect the driver's face and eyes by using Viola-Jones Algorithm that includes Haar Classifiers that showed significant advantages regarding processing time and correct detection algorithms. A prepared scenario is tested in a designed simulator that is used to simulate real driving conditions in an indoor environment. The system is added in real-vehicle and tested in an outdoor environment. Whenever the system detects the distraction or sleepiness of the driver, the driver will be alerted through a displayed message on a screen and an audible sound for more attention. The results show the accuracy of the system with a correct detection rate of 82% for indoor tests and 72.8 % for the outdoor environment.

1. Introduction

The major factor of driving on roads is the driver's attention, and once this attention is lost, major accidents could happen. According to the National Highway Traffic Safety Administration (NHTSA) [1], 153297 car crashes happened due to drowsiness in the period from 2011 to 2015. Another study by Kuwait times [2] stated that for more than 80,000 accidents, 95% of them happened due to lack of attention. The attention of the driver can be diverted through many things; using mobile phones, changing radio stations, eating and drinking, and daydreaming. In addition to that, sleepiness due to stress or fatigue; when the driver is sleepy or tired, his reaction will be slower than the normal driver which leads to accidents. There are many symptoms that can help detect sleepiness or distraction of the driver, the main symptom is the eyes of the driver. One of the ways that will assist the driver to pay attention while driving is to add an eye-tracking system that uses a camera to detect sleepiness due to stress, fatigue or any distraction. The system will alert the driver when his attention is distracted. This system is to be added to the wearable bracelet, which is used to monitor the physiological parameters of the driver [3].

2. Objectives

Real--time eye-tracking system is used to track the driver's eye. When the driver is drowsy or distracted his response time to react in different driving situations is slow. Therefore, there will be higher possibilities of accidents. There are three ways of detecting driver's drowsiness. The first one is the physiological changes in the body like pulse rate, brain signals and heart activity which can be detected by a wearable bracelet system. The second way is behavioural measures for example sudden head nods, eye closure, blinking, and yawning which is achieved by the proposed eye tracking system. The third way is vehicle based like lane position and steering wheel movements. Based on literature study conducted, the eye tracking system is the most accurate and precise way to detect drowsiness and fatigue. [4] In addition to that, it is used to detect the driver's attention on road which might happen due to texting on mobile phone, changing radio station or chatting with passengers. The paper revolved around the design of the eye tracking system. The system consists of a camera to track the driver face and detect the eyes and interactive screen for user interface with the system. Different locations are studied, and the best location is chosen and tested, the system is tested on a designed simulator. The simulator is consisting of a steering wheel and pedals as a cockpit. A robotics rover is controlled via RF signals as a vehicle in which is controlled manually using the simulator wheels and pedals. In case of eyes closing detection, the interactive screen alerts the drivers by a message and hearable sound. The robustness of the system is studied, and false detection rate is detected in in-door and out-door testing environments.

*Sherif Said, American University of the Middle East, +965 2225 1400 Ext. 2168, sherif.said@aum.edu.kw

The objective of the designed system aims the following five points:

- Affordable: The systems must be affordable as the price is one of the main factors that kept on mind during design phase.
- Portable: The systems to be portable and easy to install in different vehicles models.
- Safe: The safety of the system is achieved by choosing the appropriate location for each component.
- Fast: The response and processing time to react in case of driver's emergency is one of the keys factors since the accident happens in few seconds.
- Accurate: The system must be accurate; therefore, the most accurate algorithms have been chosen.

3. Literature Review

Distracted driving is a serious and growing threat to road safety [5] .Collisions caused by distracted driving have opened an investigation of the US Government and professional medical organizations [6] during the last years There is not an exact figure regarding statistics about accidents caused by inattention (and its subtypes) since studies are made in different places, different time frames and therefore, different conditions. The distraction and inattention account for somewhere between 25% and 75% of all crashes and near crashes.

The use of in-vehicle information systems (IVISs) trend is critical [7] due to the fact they manual, induce visual, and cognitive distraction may have an effect on performance of driving in qualitatively distinct ways. Moreover, the advancement and incidence of personal communique gadgets has exacerbated the problem during these remaining year's [8]. Some of these elements can result in the increment of the wide variety of obligations subordinate to riding pastime. Those obligations, particularly secondary tasks, which may additionally cause distraction [9], include drinking, eating, tuning the radio or the act of taking something or the use of cell phones and other technologies.

The secondary duties that take drivers' eyes off the forward roadway lessen visual test [10] and growth cognitive load can be specifically risky. For instance, the use of mobile phones even as driving, consistent with naturalistic studies [11], causes heaps of fatalities inside the US each year [12].

To tune the irradiation of IR illuminators many developments have been done recently. To let IR illuminators, operate in multiple glass reflection, distinct lighting conditions, and varying direction of gaze must be tuned. Research has been conducted to combine appearance-based methods with active IR methods. The advantage of this combination will be that, this method can do eye tracking even when the pupils are not bright due to different interferences from external illumination. Along this model, an appearance model is also incorporated with the use of shift mean tracking and vector machine support in both eye tracking and detection [13].

There are two types in which eye tracking and detection can be classified namely, Active (infrared) IR based method and method of passive appearance. A bright pupil effect is utilizing by active IR illumination method. This method is simple and effective for easy tracking and detection of eyes. The principle of working is a differential infra-red scheme [14]. This method utilizes two infrared sources of frequencies. A distinct glow in the pupil is produced when the first image is captured at 850 nm by these infrared lights. 950 nm infrared source is used by the second image for illumination that displays a dark pupil's image. The first and the second image both are synchronous with the camera and the only difference exists on the pupil region brightness. Blobs are identified after post processing the pupil and works for eyes tracking [15], Several factors affect the success rate. Pupil's size and brightness face orientation, interference of external light (lights from street lights and other vehicles) driver distance from camera. External light intensity should be limited. Another problem is the glints and reflection from glasses.

There are two steps involved in appearance-based methods which are: detecting face to extract eye region and after that detection of eye from eye windows. In order to overcome face detection method different approaches like: principal and independent components, neutral network, and method of skin colour based. There are some constraints in each method: images without expression, frontal view, short changes in conditions of light, background that is uniform and so on. To track and locate eyes of the driver Papanikolopoulos and Eriksson present a system. The approach used was a symmetry-based approach. It is used to locate face in grayscale image and after that eyes are detected and tracked.

If the eyes are closed or open, this can be determined by template matching [16]. For the identification of driver fatigue, a system of non-intrusive vision-based system was proposed by Papanikolopoulos and Singh [17]. A graphics video camera is used by the system that focus on face of driver and continuously observe eyes of driver to detect micro sleeps.

Haar like feature can be used to detect face and Michael Jones and Paul Viola proposed this method. Using integral images Haar like features will be quickly computed. AdaBoost method is used to train this algorithm. Combination of different weak classifiers are used to form a strong classifier. The advantage of this technique is that, the combination will let detector work in cascaded manner and the first stage classifier faces like regions will be more intensively processed. The rate of detection with this method was above 95%. In a few processes authors used colour, facet, and binary records to come across eye pair candidate regions from input photograph, then extract face candidate location with the detected eye pair. SVM (Support Vector Machines) are used to detect region of candidate face and eye pair [18] .There are approaches in which eye's dimensions and appearance are used. Mostly tracking and detection of eye's pupil or iris is required by eye tracking applications. Iris and pupil can be modelled by five shape parameters, as they both appears to be elliptical depending on the viewing angle. In order to detect iris or pupil, Hough transform is the effective way, but it requires detection of explicit features.

There are four stages in Viola-Jones algorithm, consisting of integral image, Haar features, cascade and AdaBoost [19] .To begin with, Haar features detection can be summarized in three steps. First, convert the colored, RGB, images to grey-scale images. After converting the image to grey-scale, integral images

are generated. The integral image is a summation of the pixel values of the original image. The summation of the pixels are calculated by choosing a coordinate (x,y) and summing all the values to the top-left of the point including the point itself. Second, from different levels of integral images the Haar wavelets are obtained. Haar-like features are a combination of two or more rectangles that have different contrast [20]; they can be two-rectangle, three-rectangle, or four-rectangle, and so on. The third step is the AdaBoost, which is an algorithm that looks at relevant features and is used to detect objects, in this case the face and eye. When the features are extracted, a stage comparator will sum all the features and compares them with the threshold. After these steps are done, there is the cascade step; and in this step the stages are cascaded to eliminate any non-face candidates to improve the computation time an accuracy of the detection.

One of the algorithms that were used to detect the pupil of the driver's eye is the Circular Hough Transform [21]. This algorithm detects any circular object in the images, in these cases the pupil of the eyes. The first step is defining the radius or the average radius of the human's pupil, two circle equations are needed, one for each eye. The algorithm will first filter the image to reduce the noise, and then the image will be converted into a grey-scale image. The edge of the eyelids will be detected to look at the circles between the two eye lids, which are the pupils.

A correlation matching algorithm, which is used only on raw images, is used to identify the targeted position, which is the eye, and then tracking the motion of the detected eyes. The reason for using only raw images is because all the information of the images is kept, therefore, making the process easier. First, there is a source image which is the original image of the driver. Second, the template image is the image that will be compared to the source image. To compare the source and template images, an equation is used where the matrices of the two images are compared together to find the similarities between them by using the Normalized correlation equation

Another method that is used and based on Digital Image Processing algorithms is the opened and closed eye template. The eyes of the driver are continuously scanned to determine whether the driver is drowsy or not. If three consequent frames were captured and the eyes were found to be closed, the system will state the driver is drowsy.

In [22], the author that uses DIP algorithms and continuous eyes monitoring to detect drowsiness of driver. The best indicator of fatigue state is the Micro sleeps as they are for the short period of almost 2 to 3 seconds of sleep. This presented system contains two cameras namely: narrow and wide angle camera. Narrow angle camera focuses on the eyes and monitors gaze and eyelid movements, while wide angle camera focuses on face and monitors facial expression and head movement. If some abnormal actions are detected by the system, it gives alarm warnings and reduces the speed of the system. This system not only detect the drowsiness of the driver but also detect the distance of the car from objects or other cars using Ultrasonic sensor. When these sensors detect objects they warn the driver and reduces speed.

The above algorithms are implemented using several software's including GUI (MATLAB), and Python (OpenCV). So briefly, the algorithms will work as follow: the camera will capture an image, after that, the algorithm will scale and extract features from the image, and then the algorithm will classify features.

Many driver-monitoring systems have been developed while working on driver fatigue detection. All these systems focus on providing information to drivers that will facilitate their driving and increase traffic safety. These can be divided into systems, DAISY (Driver Assisting System) and DAS (Driver Assistance System. Driver assistant system DAISY is used in German motorways to warn and monitor driver for lateral and longitudinal control [23] .The second system DAS is developed by a group at Australian National University [24][25]. To monitor the driver, a dashboard mounted face LAB head and eye tracking system is used. Driver performance is detected by algorithm known as distillation algorithm. Driver is provided with the feedback on deviation in lane tracking by using steering wheel force feedback that is related to offset lateral estimation by the tracker of lane.

4. Methodology

The driver's distraction and sleepiness while driving can be detected by adding a camera on the vehicle dashboard to provide a real-time tracking of the driver's face. In that case, the task of detecting the driver's drowsiness, fatigue and distraction will be much easy and more accurate.

After the literature study about the systems, algorithms and methods used for eye-tracking system. Some important factors were noticed that help in the development of the proposed system in this paper. The average response time between the analysis and the alert process to be 50 milliseconds. The alarm should be loud enough to alert the driver. The system should operate even when the driver is using sunglasses and at night without flash to ensure non-intrusiveness. Different methods of detection are mentioned in Table 1 Methods of eye tracking systems.

Table 1. Different systems with accuracy

Method	TN	FP	FN	TP	Overall Accuracy
Blinking	91.72%	8.18%	10.95%	89.05%	90.74%
Lateral Position	89.39%	10.61%	20.95%	79.05%	85.37%
Steering Angle	88.48%	11.52%	14.77%	85.23%	87.22%

Where:

TN: Eyes closed system, detected closed
FP: Eyes closed, system detected open
FN: Eyes open, system detected closed
TP: Eyes open, system detected open

This means that available systems are most accurate in eye tracking with detection of blinking rate, and the least accurate method was the lateral position. Therefore, based on these results, the best method to detect drowsiness is by eye-detection.

To sum up, the systems were able to detect sleepiness and distraction for almost every eye-type and for both genders. Also, the most accurate method to detect driver's distraction is through

face and eye-detection. By using software, these researches were tested then implemented successfully, and by using the display screen and the buzzer, the driver is alerted and this will help in reducing the number of car accidents and increase the safety on roads.

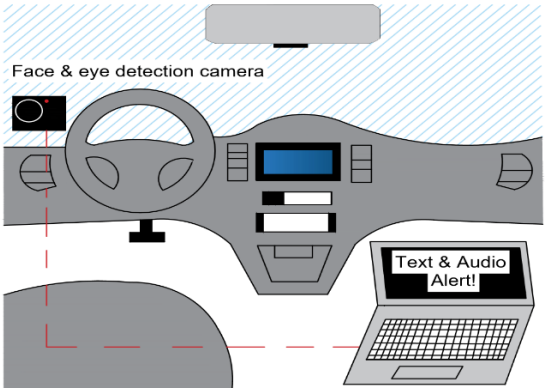


Figure 1. Proposed Eye-tracking System Sketch

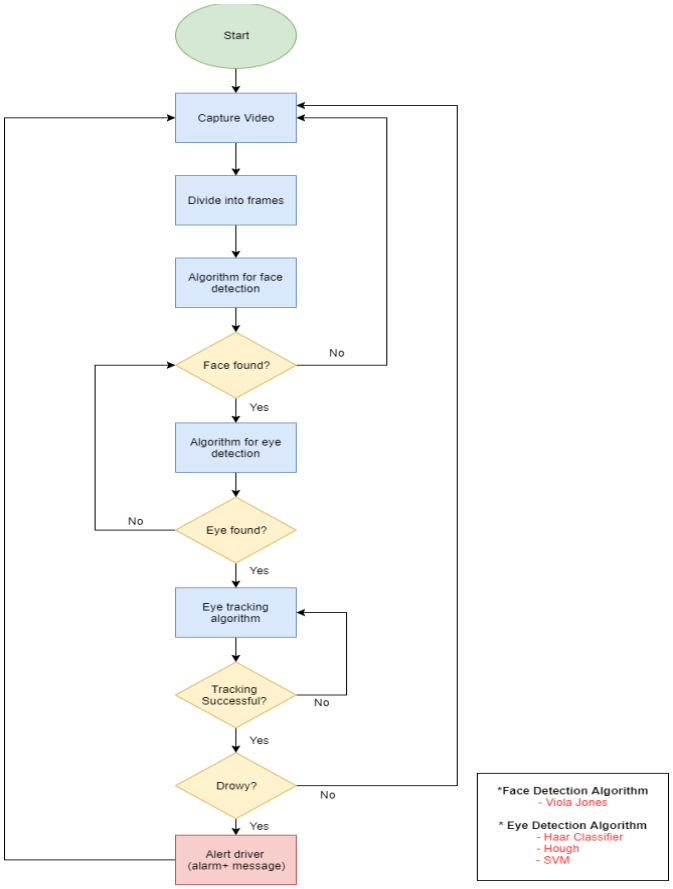


Figure 2. Flow-chart of the eye-tracking process

5. System Components

The proposed system consists of a high-resolution camera, an LCD, a speaker, and a microprocessor. The screen to be placed on the dashboard to the right-hand side of the driver. The camera will be positioned on the dashboard to the left-hand side of the driver and will be fixed by a camera holder. To start the system, the camera will be continuously recording a video of the driver and this video will be analyzed on PC initially then replaced by microprocessor later. The analysis will be done through many algorithms to detect the eyes of the driver. When the driver seems drowsy or distracted, the system will show an alert on the display screen among with a sound alert. Figure 2 shows the proposed design in a form of sketch for the system. The system flow chart that need to be implemented is explained in the flow chart. The real-time camera will stream an online video. An eye-tracking algorithm can be used to extract the features of the eyes. The classifier will play a role in the detection of open/close of the eye.

6. Design Specifications

The design of the real-time eye tracking is implemented to be an added-on solution that can fit in any vehicle. The purpose of the system is to alert the driver in case of drowsiness and distraction. The system will alert the driver as precaution event. In emergency cases, the system should interact with the vehicle and take over the control. This system is an added value to the wearable bracelet that couldn't detect the sleepiness and distraction.

6.1. High Level Design

A simulator cockpit is designed and implemented in a professional way for testing the system. A steering wheel among with pedals are developed in an ergonomic way to simulate driving scenarios on road.

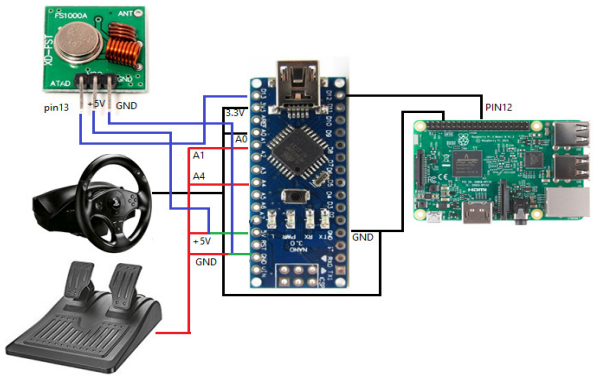


Figure 3. Connection of Raspberry Pi and Arduino Nano (Steering wheel)

6.2. Low Level Design

For the Low-Level design, the connections are made with Arduino and Raspberry Pi (see **Error! Reference source not found.**). The Raspberry Pi pin 12 is connected to the input pin 12 of the Arduino Nano that is attached to the steering wheel. In addition, the RF transmitter, the steering wheel and the pedals are all connected to the Arduino Nano as shown. The RF communication module is used to overcome the delay problem that happened when the control signals sent via Bluetooth. An Arduino Uno is attached to the rover (see **Error! Reference source not found.**), which shows the connection of the RF receiver, the LED, used to show that the rover is moving in autonomous mode, and the push button that the user must press to return to manual driving.

7. System Hardware Specifications

7.1. Camera

The Camera used in detection of the eye open/close is Logitech C290. The specifications of the camera are presented in

Table 2. The most important factor in selecting the camera is the number of frames that the camera can capture in one second. For this camera is 30 frames per second (fps).

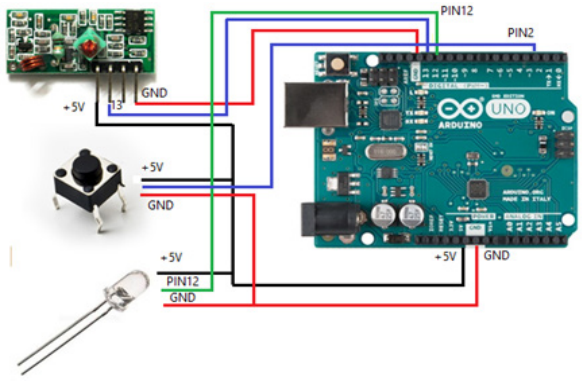


Figure 4. Connections of Arduino Uno (Rover)

Table 2. Specifications of the Camera

Device Type	Web camera
Connectivity Technology	Wired
Digital Video Format	Colour
Max Digital Video resolution	H.264
Features	1080p Full HD movie recording,
Battery	None
Audio support	Yes
Computer interface	USB 2.0
OS required	Microsoft Windows 7, Microsoft Windows Vista, Microsoft Windows XP SP3 or later
Connector type	4 pin USB Type A

Table 3. Raspberry Pi3 Specs

SoC	BCM2837
CPU	Quad Cortex A53 @ 1.2GHZ
Instruction set	ARMv8-A
GPU	400MHz VideoCore IV
RAM	1 GB SDRAM
Storage	Micro-SD
Ethernet	10/100
Wireless	802.11n/ Bluetooth 4.0
Video Output	HDMI/ Composite
Audio Output	HDMI / Headphone
GPIO	40

7.2. Raspberry Pi 3

Raspberry Pi 3 Is a tiny credit card size computer. Just add a keyboard, mouse, display, power supply, micro SD card with installed Linux Distribution and you'll have a fully-fledged computer that can run applications from word processors and spreadsheets to games.

7.3. Display & Audio 1280x800 IPS Screen

This screen has excellent resolution (1280x800) and IPS display so it is bright, crisp and looks good from any angle. This Screen even has HDMI audio support and can drive two 4-ohm speakers directly.

Table 4. IPS Screen Specs

Dimensions of Screen	105mm x 160mm x 3mm / 4.1" x 6.3" x 0.1"
Weight of Screen	91g
Power of Screen	9 VDC
Display Ratio	16:10
Resolution	1200 x 800
Visible Area	150mm x 95mm 16:10
Display Dimensions	162mm x 104mm x 4mm (6.4" x 4.1" x 0.2")
Brightness	400cd/m2
Contrast	800:1
Display	HSD070PWW1
Weight	290g/10.2oz
HDCP	None

8. Image Processing

For implementation of the system. A simulation of the eye tracking system is done using PC. The software used is OpenCV. For face and eye detection using Viola Jones and Haar Cascade Classifier algorithms were applied. The idea is to detect eyes are open or closed. When the eyes were closed, the system attempted to alert by showing a display message and sound. After the proof of concept, Raspberry Pi board as a main Processor for the system replaces the PC.

The basic idea of detecting the state of the driver revolves around the detection of the eyes. The recent advancement in the image-processing field, there are now multiple real-time methods that could enable the detection and tracking of multiple objects. Bag-of-Words models, Histogram-of-oriented gradients (HOG), Deformable Parts Models, Exemplar models and Viola-Jones are some of the widely used methods for object detection.

Viola-Jones method is one of the most widely used method of object detection. The main feature that makes this method so popular is its ability training is slow but detect fast. Haar basis feature filters are used in this algorithm avoiding, thus avoiding multiplications.

Detection happens inside a detection window. The minimum and maximum window size is selected, and for each size a sliding

step size is selected. The detection window moves across the selected image as follows:

Cascade -connected classifiers contained inside each face recognition filter. The classifier takes care of scanning at a rectangular subset of the detection window and recognize if it looks like a face or not. The next classifier is applied in case the face detected on the off chance that all classifiers give a positive answer, at that point the filter gives a positive answer and the face is perceived. If nothing detected, the next filter in the set of N filters should execute.

Each classifier is composed of Haar feature extractors (weak classifiers). Each Haar feature is the weighted sum of 2-D integrals of small rectangular areas attached to each other. The weights may take values ±1. Haar feature extractors are scaled with respect to the detection window size.

Viola-Jones algorithm consists of four stages, Haar features, integral image, AdaBoost, and cascade. To begin with, Haar features detection can be summarized in three steps. First, convert the colored, RGB, images to grey-scale images. Then, integral images are generated. The integral image is a summation of the pixel values of the original image. The summation of the pixels are calculated by choosing a coordinate (x,y) and summing all the values to the top-left of the point including the point itself. Second, different levels of integral images are used to obtain the Haar wavelets. Haar-like features are a combination of two or more rectangles that have different contrast; they can be two-rectangle, three-rectangle, or four-rectangle, and so on. The third step is the AdaBoost, which is an algorithm that looks at relevant features and is used to detect objects, in this case the face and eye. When the features are extracted, a stage comparator will sum all the features and compares them with the threshold. After these steps are done, there is the cascade step; and in this step the stages are cascaded to eliminate any non-face candidates to improve the computation time an accuracy of the detection.

8.1. Algorithm Implementation

The Algorithm is applied sequential. The first step is to detect the face. If no faces detected, this means that the driver is looking to the right or left or down. A counter start counting, and an alert message is sent after 1 second. If faces detected, then the algorithm will extract the eyes from the face. If the driver's eyes opened don't count any values. In case the driver closed his eyes then a counter start counting based on the required timing. After that time an alert message is sent to driver's. Figure 5 shows the application of the algorithm as a simulation of a detected eye.

9. Results

Real-time Eye tracking came up as an added-solution to the wearable biosensors bracelet. This solution helps in an accurate real-time monitoring of the driver's eyes while driving. This system aims to assist the driver not only in case of drowsiness and in case of Fatigue; it can assist in case of driver's distraction due to changing radio stations, chatting with passengers or texting while driving.

The system implemented as a cockpit for testing in-door to validate the system. The system is installed and tested in a real-vehicle to proof the correctness of the proposed system.

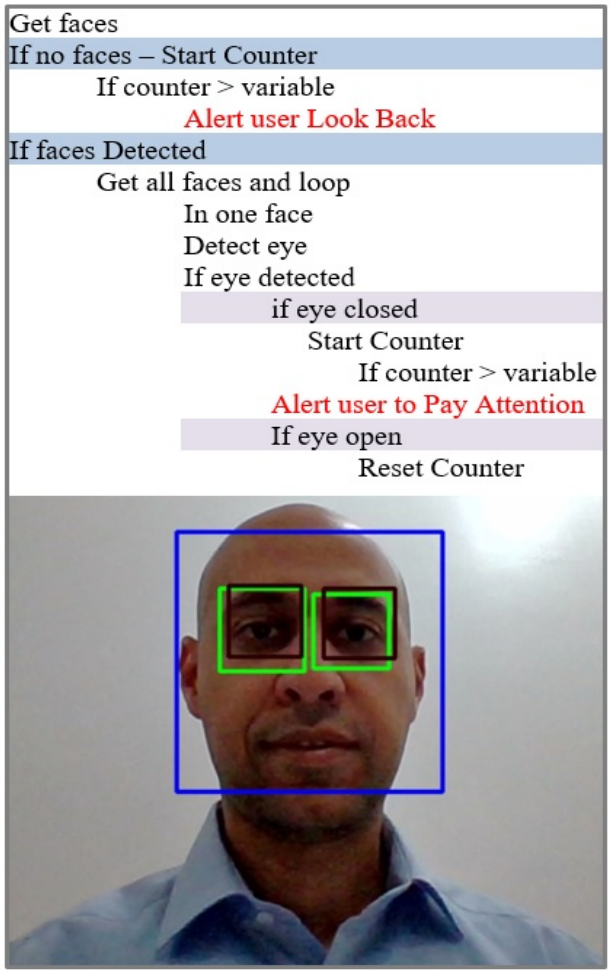


Figure 5. Eye Detection using Viola-Jones Algorithm

9.1. Indoor-System Test

The High-Level design of the simulator shown below (see **Error! Reference source not found.**), where the driver is seated in front of the steering wheel. The steering wheel gives the subject the feeling of controlling a vehicle. The driver is facing a camera and the display screen. In addition, there are pedals that control the rover when the driving mode is manual. One pedal is used to move the car forward and the other one moves it backwards.

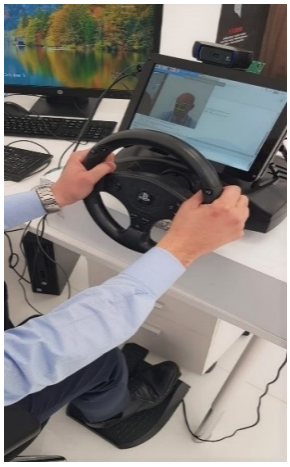


Figure 6. Simulator Cockpit Design

The system tested in-door on 20 persons to record a small data-set for analysis and validation of the system. The system will be coupled to the prototype explained earlier, where the driver sat down in front of the steering wheel and imitated distraction or sleepiness as shown in figure. 6. The subject is controlling the robotics rover as he/she is driving a vehicle on roads. In case the user is distracted for 1.5 seconds (pre-set value), the system will respond by a message on the screen Look back! Moreover, the system reacts by switching the rover mode to autonomous in which the rover will run obstacle avoidance algorithm as a simulation of cars and obstacles on roads. In case, the subject closed his/her eyes for 1 second (pre-set value), the system will react by switching the rover mode to autonomous mode.

The recorded data of the subjects for in-door testing is shown in Table 5 and the graph is shown below (see **Error! Reference source not found.**). The results of average correct detection rate are 82 %. This result is very promising to test the system in real vehicle.

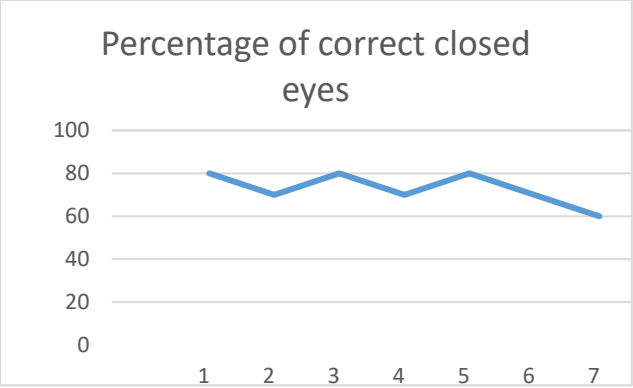


Figure 7. Correct Detection Percentage

Table 5. In-door Real-time Eye tracking Recorded Data

Index	Age	Gender	Samples	Correct Detections	False Detections	Percentage
1	20	M	10	8	2	80
2	23	M	10	7	3	70
3	26	M	10	8	2	80
4	21	M	10	9	1	90
5	36	M	10	10	0	100
6	25	M	10	7	3	70
7	45	M	10	7	3	70
8	36	M	10	8	2	80
9	24	M	10	9	1	90
10	62	M	10	10	0	100
11	45	F	10	8	2	80
12	52	F	10	7	3	70
13	42	F	10	9	1	90
14	35	F	10	10	0	100
15	26	F	10	7	3	70
16	29	F	10	8	2	80
17	27	F	10	7	3	70
18	28	F	10	8	2	80
19	39	F	10	7	3	70
20	34	F	10	10	0	100
				Average Correctness		82

9.2. Out-door System Test

The system is installed in two different car models for real driving conditions testing as shown in below (see **Error! Reference source not found.**). Once, the system detects eye-closing for 1 second (pre-set value) or distraction detection for 1.5 second (pre-set value) an audible alert sound among with alert message sent to the user for driver's assistance on road.

Figure 8. Installation of the System in real-vehicles

The system tested on seven subjects. The testing scenario is to drive the car at speed around 40-50 Km/hr. The driver will act based on the assistant instructions. The assistant task is to direct the driver to close eye or distract the eyes by looking right/left and record the data of correct detection to validate the algorithm and system on roads. The recorded data is shown in **Error! Reference source not found.** and the results of average correct detection rate of closed eyes is 72.8 % with an average correct detection rate of distracted eyes of 67.14 % as shown below (see **Error! Reference source not found.**).

As noticed, the results of correct detection reduced compared to the in-door test results. The reasons of this decrease are that the car on the streets is exposed to different lighting conditions that needs a pre-setting. Also, the driver movements in the real-driving conditions increases which needs higher frame rate of camera.

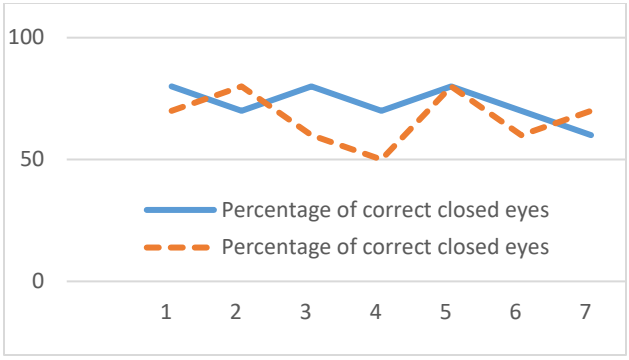


Figure 9 Correct Detection of eye closed and distracted eyes in out-door testing environment

Table 6. Out-door Real-time Eye tracking Recorded Data

Index	Age	Gender	Samples	Correct Detections of closed eyes	Correct Detections of Distracted eyes	Percentage of correct closed eyes	Percentage of correct distracted eyes
1	30	M	10	8	7	80	70
2	25	M	10	7	8	70	80
3	21	F	10	8	6	80	60
4	22	F	10	7	5	70	50
5	24	F	10	8	8	80	80
6	20	F	10	7	6	70	60
7	22	F	10	6	7	60	70
				Average Correctness		72.857143	67.142857

The system testing out-door is shown below (see **Error! Reference source not found.**). The system is tested with the PC to record the system test in out-door environments. The system is validated, and the results were promising to improve the algorithm for more scenarios.

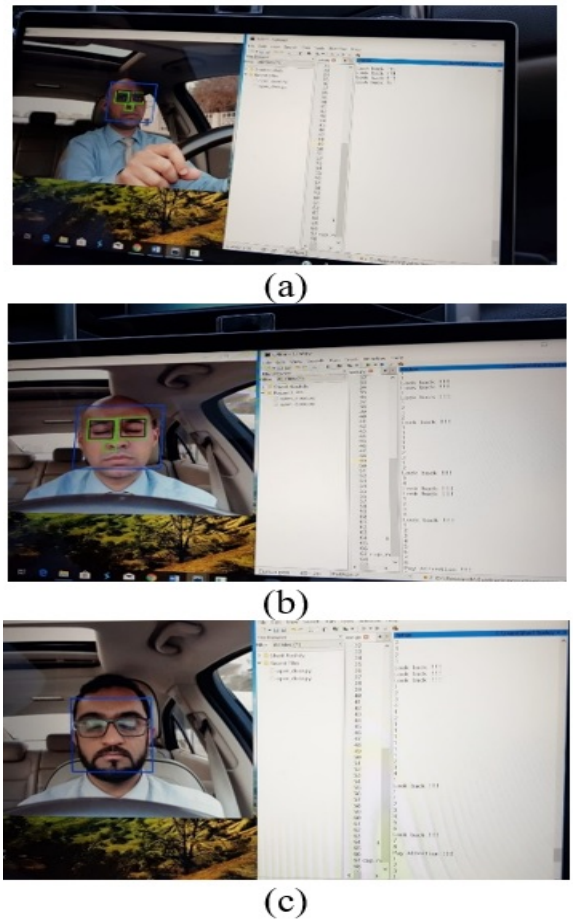


Figure 10. (a) Distracted driver, (b) closed eyes driver, (c) closed eyes driver

10. Conclusion

The complete system is implemented with a developed simulator cockpit for in-door system validation and data collection purposes. The measured average correct detection of the system is 82 %. This system is tested in subjects in average of 33 years old and showed no issues regarding the age. It is valid for people wearing eyeglasses without any problem. In out-door environment, the results of average correct detection rate of closed eyes are 72.8 % with an average correct detection rate of distracted eyes of 67.14 %.

11. Future Works

- The system can't detect during night due to lack of camera version
- The response time of the system when the driver closes his/her, eyes should be a function of vehicle speed.
- The system should be adapted to the looking of mirrors scenario.
- Sensor fusion of the camera with all the wearable biosensors.
- Use Night-vision camera to detect the eyes at night.
- Use better specifications camera will increase the efficiency of the system

References

[1] https://www.nhtsa.gov/, 4/2017.

[2] "Kuwait Times," 15 10 2017. [Online]. Available: http://news.kuwaittimes.net/website/.
[3] Sherif Said ; Samer AlKork ; Taha Beyrouthy ; M Fayek Abdrabbo, "Wearable bio-sensors bracelet for driveras health emergency detection," in Biosmart ,IEEE, Paris, France, 2017.
[4] Singh Himani parmar ; Mehul Jajal ; Yadav priyanka Brijbhan, "Drowsy Driver Warning System Using Image Processing," Nternational Journal Of Engineering Development And Research , 2017.
[5] "http://www.who.int," 2016. [Online].
[6] Llerena, L.E.; Aronow, K.V.; Macleod, J.; Bard, M.; Salzman, S.; Greene, W.; Haider, A.; Schupper, A. J. , " An evidence-based review: Distracted driver.," Trauma Acute Care Surg., vol. 78, p. 147–152, 2015.
[7] Bennakhi, A.; Safar, M. , "Ambient Technology in Vehicles: The Benefits and Risks," Procedia Comput. Sci., vol. 83, pp. 1065-1063, 2016.
[8] Liu, T.; Yang, Y.; Huang, G.B.; Lin, Z, "Detection of Drivers' Distraction Using Semi-Supervised Extreme Learning Machine," in In Proceedings of ELM-2014; Springer, Berlin, Germany, 2015.
[9] Simons-Morton, B.G.; Guo, F.; Klauer, S.G.; Ehsani, J.P.; Pradhan, A.K, "Keep your eyes on the road: Young driver crash risk increases according to duration of distraction," J. Adolesc. Health, vol. 54, p. S61–S67, 2014.
[10] Recarte, M.A.; Nunes, L.M., " Mental workload while driving: Effects on visual search, discrimination, and decision making," J. Exp. Psychol. Appl. , vol. 9, pp. 119-137, 2003.
[11] Klauer, S.G.; Guo, F.; Simons-Morton, B.G.; Ouimet, M.C.; Lee, S.E.; Dingus, T.A, "Distracted driving and risk of road crashes among novice and experienced drivers," N. Engl. J. Med., vol. 370, p. 54–59, 2014.
[12] Bergmark, R.W.; Gliklich, E.; Guo, R.; Gliklich, R.E. , "Texting while driving: The development and validation of the distracted driving survey and risk score among young adults.," Inj. Epidemiol., 2016.
[13] Xia Liu, "Real-time eye detection and tracking for driver observation under various light conditions," Intelligent Vehicle Symposim, 2002.
[14] R. Grace, "Drowsy driver monitor and warning system," in Proc. Int. Driving Symp. Human Factors in Driver Assessment, Training andVehicle Design, 2001.
[15] Z. G. Yuan, "A real-time eye detection system based on the active IR illumination".
[16] Eriksson, and N.P. Papanikotopoulos M., "Eye-tracking for detection of driver fatigue," Proc .Int. Conf. Intelligent Transportation Systems, pp. 314-318, 1997.
[17] S. Singh, and N.P. Papanikolopoulo S., "Monitoring driver fatigue using facial analysis techniques," in Proc. Int. Conf. Intelligent Transportation Systems, Tokyo, 1999.
[18] Hyungkeun Jee, Kyunghee Lee, and Sungbum Pan, "Eye and Face Detection using SVM," in Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference, Melbourne, Australia, 2005.
[19] Jay D. Fuletra ; Viral Parmar, "Intelligent Alarm System for Dozing Driver using Hough transformation," IJEDR, vol. 2, no. 2, pp. 2797-2800, 2014.
[20] Anirban Dasgupta, Anjith George, S. L. Happy, and Aurobinda Routray, "A Vision-Based System for Monitoring the Loss of Attention in Automotive Drivers," IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, vol. 14, no. 4, 2013.
[21] Hari Singh ; Jaswinder Singh, "Human Eye Tracking and Related Issues: A Review," International Journal of Scientific and Research Publications, vol. 2, no. 9, 2012.
[22] Mitharwal Surendra Singh L. ; Ajgar Bhavana G. ; Shinde Pooja S. ; Maske Ashish M. , "EYE TRACKING BASED DRIVER DROWSINESS MONITORING AND WARNING SYSTEM," International Journal of Technical Research and Applications, vol. 3, no. 3, pp. 190-194, 2014.
[23] R. Onken, "DAISY: an adaptive knowledge-based driver monitoring and warning system," in Proc. Vehicle Navigation and Information Systems Conf., 1994.
[24] L. Fletcher, N. Apostoloff, L. petersson, and A. Zelinsky, "Vision in and out of Vehicles," IEEE Trans. Intelligent Transportation Systems, pp. 12-17, 2003.
[25] F Sayegh, F Fadhli, F Karam, M BoAbbas, F Mahmeed, JA Korbane, S AlKork, T Beyrouthy "A wearable rehabilitation device for paralysis," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp.1-4.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 469-478 (2018)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

CNN-based Automatic Coating Inspection System

Lili Liu[1], Estee Tan[1], Zhi Qiang Cai[*,2], Xi Jiang Yin[1], Yongda Zhen[*,1]

[1]*Department for Technology, Innovation and Enterprise, Singapore Polytechnic, Singapore, 139651*

[2]*School of Electrical and Electronic Engineering, Singapore Polytechnic, Singapore, 139651*

ARTICLEINFO

Article history:
Received: 30 November, 2018
Accepted: 13 December, 2018
Online: 26 December, 2018

Keywords:
Coating corrosion assessment
Deep transfer learning
Instance aware segmentation
Hidden corrosion assessment
Active thermography

ABSTRACT

The application of protective coatings is the primary method of protecting marine and offshore structures from corrosion. Coating breakdown and corrosion (CBC) assessment is a major aspect of coating failure management. Evaluation methods can result in unnecessary maintenance costs and a higher risk of failure. To achieve a comprehensive collection of data for CBC assessment, an unmanned arial system (UAS), assisted by the latest technological innovations, will be used to facilitate data collection in inaccessible locations. A convolutional neural network (CNN)-based CBC assessment system is developed to provide objective assessment of the severity of coating failure. This method is more suitable for inspecting large areas by capturing and analyzing pictures/videos of the target area than the surveyor's existing manual inspection solution. In this paper, deep learning-based object detection in the CBC assessment system has been developed to provide an effective CBC assessment for the marine and offshore industries. By using active thermal imaging, it can identify corrosion behind the coating. This will greatly improve the efficiency and reliability of coating inspection.

1 Introduction

Protective coatings are the main route used to prevent corrosion of marine and offshore structures. Checking the protective coating is a key issue for asset management, but traditional visual inspection methods are time-consuming and labor-intensive. The project aims to develop an automatic coating inspection system for corrosion management applications. The developed system will be able to quickly and comprehensively screen and evaluate coating conditions and can be used as a scanning tool to help investigate the identification and classification of coating failures.

In order to achieve a comprehensive collection of data for coating condition assessment, a micro-aerial vehicle (MAV) assisted by the latest technological innovations, will be used to facilitate data collection in inaccessible locations. An image-based coating breakdown and corrosion (CBC) assessment system has been developed to provide objective assessment of the severity of coating failure. After inspection is completed, the automatic CBC assessment system (A-CAS) is able to generate an inspection report. The inspection method developed and the test data obtained in this paper can assist the inspection and maintenance team to improve efficiency and productivity. It reduces maintenance cost, improves structural integrity and minimizes risk of failure. Compared to conventional inspection, A-CAS can reduce coating failure evaluation time and manpower.

The structure of the paper is as follows. Section 2 lists related works, Section 3, Section 4 and Section 5 present innovative detection solutions for CBC evaluation, introduces image-based CBC evaluation methods, and explains non-destructive hidden corrosion assessment method. Finally, conclusions are drawn in Section 6.

2 Background

This paper is an extension of work originally presented in the 13th IEEE conference on industrial electronics

[*]Zhi Qiang Cai, School of Electrical and Electronic Engineering, Singapore Polytechnic, Singapore, 139651, 65 - 6772 1542 & cai_zhi_qiang@sp.edu.sg

[*]Yongda Zhen, Department for Technology, Innovation and Enterprise, Singapore Polytechnic, Singapore, 139651, 65 - 6772 1455 & zhen_yongda@sp.edu.sg

https://dx.doi.org/10.25046/aj030655

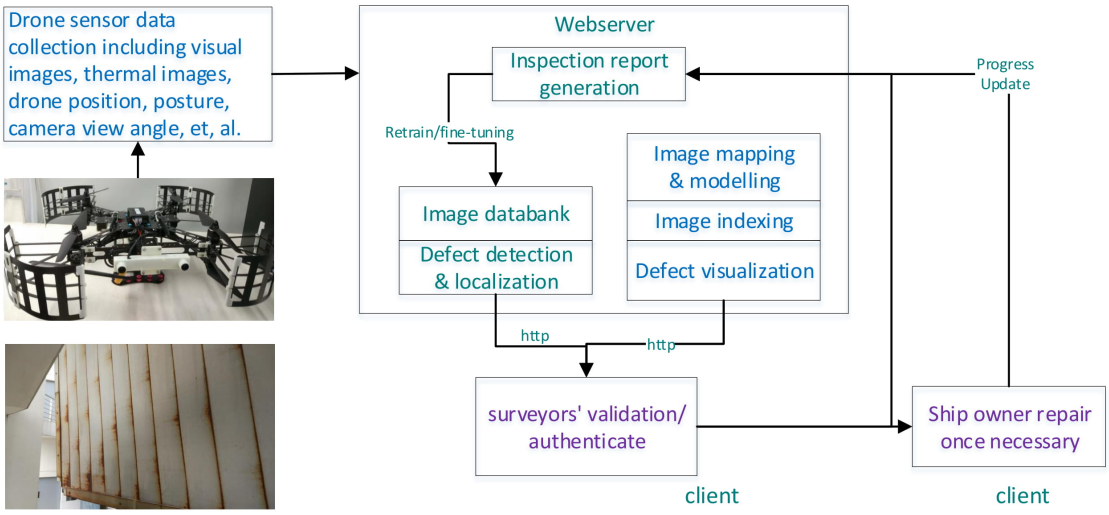


Figure 1: UAS for automated CBC assessment

and applications (ICIEA), 2018 [1]. Referring to automated vision-based coating corrosion detection, to the best of the authors' knowledge, In [2], Jahanshahi and Masri utilize color wavelet-based texture analysis for corrosion detection. Ji et al. [3] apply watershed transform over the gradient of gray images. Siegel et al. [4] choose wavelets for characterizing and detecting corrosion texture in airplanes, Zaidan et al. [5] focus on corrosion texture using standard deviation and entropy as discriminating features. Last but not the least, Ortiz et al. [6] present a solution for coating breakdown and corrosion detection which adopts a semi-autonomous MAV, and an artificial neural network (ANN) to discriminate between pixels suspected/not suspected of corresponding to coating breakdown and corrosion areas through sufficient color and texture descriptors.

2.1 UAS Facilitated Data Collection

Nowadays, visual inspection is widely used in both industry and daily life for security or maintenance checking purposes. Sometimes, visual inspections take place in areas that put the inspector in highly risky situations, hence extra precautions have to be taken, resulting in increasing inspection costs and turnaround time. All these negative factors create difficulties in inspection and would delay the subsequent processes. To meet the needs of inspection, different types of unmanned vehicles were suggested and introduced to replace manual inspection. In coming years, as the technology in aircraft flight control and video processing and transmission systems becomes mature, unmanned arial system (UAS) would be a solution. It is imperative to use MAVs to replace in-situ human inspection in order to save the cost of preparing the vessel for inspection. Indoor vessel inspection requires the aerial platform to be capable of stable, low speed flight and have anti-collision functions. For research purposes, we customized a MAV with a protective cage and a RGB-D camera for indoor anti-collision coating failure inspection to facilitate our data collection.

In this study, we propose a UAS facilitated CBC assessment system as shown in Figure 1. First of all, a customized autonomous AeroLion drone equipped with necessary inspection tools conducts vessel condition inspection; semi-autonomous navigation technology fulfills the flight capability requirement in indoor environments. The data collected from the drone are then transferred to web-server for post-image analysis including CBC detection and localization in the reconstructed 3D model. Image mapping and image indexing technologies are being developed for traceability of the inspection. The customized Aerolion drone is integrated with a Nvidia Jetson TX2 embedded processor and a ZED RGB-D stereo camera for visual-based simultaneous localization and mapping (vSLAM) mapping, navigation and data collection. Due to the drone's computational power limitation, post-image processing on a separate platform is required for detailed CBC assessment. Figure 1 visualizes an integrated post-processing system. First, a web-server will integrate CBC analysis, image mapping and the CBC visualization model. Then, the HTTP-based client application can be accessed through mobile tablets. Finally, AI-based autonomous CBC inspection results are sent to the surveyor for result verification and report generation. In this paper, we will focus on the automated CBC assessment system.

Instead of image data analysis by certified surveyors, this paper presents one deep learning based visual CBC detection system. The system can automatically recognize and analyze the coating condition by classifying and determining the type of CBC. Corrosion includes but is not limited to CBC on edges, CBC on welds, and CBC on surfaces including hard rust and pitting. Due to the limited size of the available dataset for system development, a transfer learning network was applied to learn the model for CBC detection. An AI-assisted CBC assessment system was successfully developed for the maritime industry. In addition to the visual camera, an infrared/thermal camera will be used for inspection of corrosion behind coating. Ther-

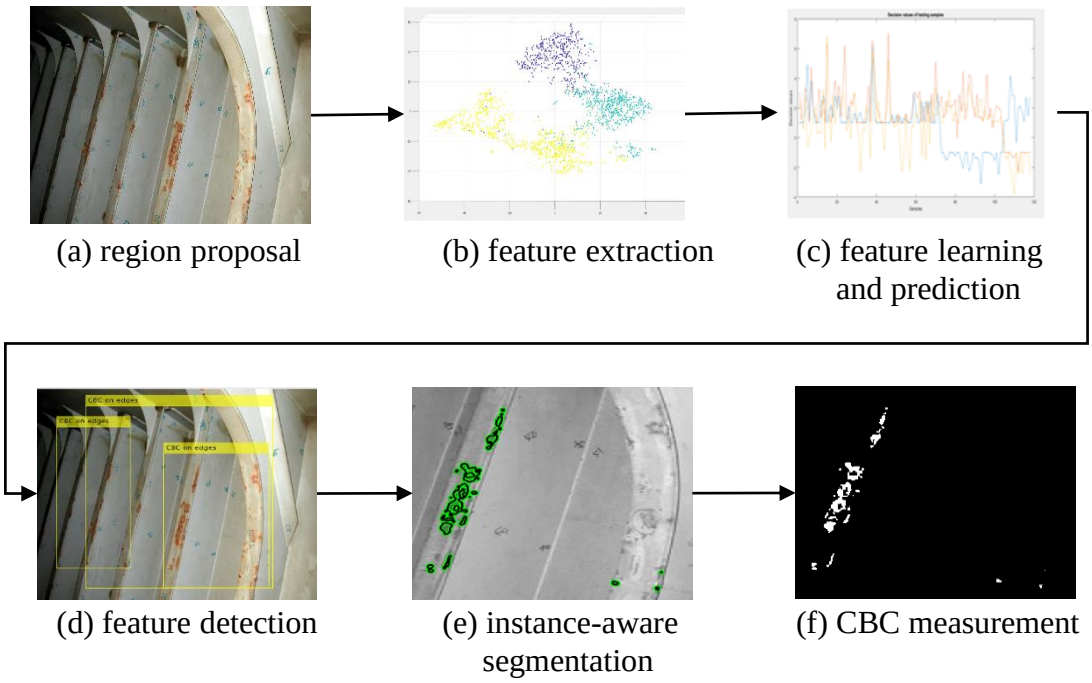


Figure 2: A-CAS system process

mal radiation from areas of the ship's surface with CBC will be different from that without CBC, and this difference can be identified and visualized in some cases. A technology of active infrared thermography to detect the hidden CBC in the vessel structures by combining an external heating source and infrared thermography is developed during the study. The thermal images collected by the drone will be also analyzed by the CBC detection system.

3 Autonomous CBC Assessment System

A vision-based protective coating inspection system was developed for automatic CBC assessment. With this program, even a relatively inexperienced surveyor will be able to make an objective judgment on the severity of CBC. This would improve the efficiency and reliability of coating inspection while reducing the time and manpower cost required. Algorithm development is divided into five phases. They are: region of interest (ROI) proposals, CBC feature extraction, CBC feature learning and prediction, CBC detection and CBC measurements.

In this paper, CBC detection and segmentation are combined for CBC instance aware segmentation. Instance aware segmentation is challenging because it requires proper detection of all CBCs in the image. In practice, due to the limited data set, transfer learning on convolutional activation feature (TLCAF) network [7] is used for CBC feature extraction and feature learning. Models are generated for CBC prediction, segmentation and measurements.

As shown in Figure 2 (b), labeled features are visualized in 3D space and the three categories including surface CBC, non-coating-failure and edge CBC can be differentiated in hyperplane. Figure 2 shows a typical example of the instance-aware semantic segmentation algorithm for coating failure assessment to speed up the inspection process. The predicted CBC ROIs are reconstructed by background removal. The ROIs are then sent to the next step for active segmentation [8] and CBC measurement. Figure 2 (e, f) demonstrates edge detection and active segmentation, where the background has been masked by active segmentation. The CBC assessment results are more accurate and constant than the ground truth images indicated in [9].

3.1 Feature Extraction and Visualization

Figure 3 lists the five classes of CBC. Different ROIs are manually labeled as different categories. The three categories used in this algorithm are surface CBC, edge CBC and non-coating-failure. Surface CBC can be further categorized into hard rust and pitting. Edge CBC can be divided into CBC on edges and CBC on welds. In order to balance dataset, the three major categories CBC including surface CBC, edge CBC and non-coating-failure are used for prediction for this work.

During the study, TLCAF network [7] is used to learn the region of interest for different types of coating failure including surface CBC and edge CBC. Bounding box is used for region of interest proposal, Faster RCNN [10, 11] framework and vgg19 model [10] are used for convolutional neural network (CNN) feature extraction, and a linear classifier is used for feature classification. Then, the ROIs are reconstructed for different types of coating failure for instance-aware semantic segmentation. Here, active segmentation[8] is used for background removal. The seeds generated

Figure 3: CBC classification

in the previous step are used for auto-guided active selective propagation. Ten iterations of propagation can generate the optimal result which matches with the ground truth which is the value provided by human surveyors' assessment.

2000 randomly selected CBC image features were chosen for the t-distributed stochastic neighbor embedding (t-SNE) [12] high-dimension feature visualization experiment. Figure 4 represents the distribution of three types of CBC features through t-SNE. High-dimension CNN features of the three types of CBC including surface-based CBC (yellow), edge-based CBC (green) and non-coating-failure (blue) can be easily distinguished in hyper-plane for feature prediction.

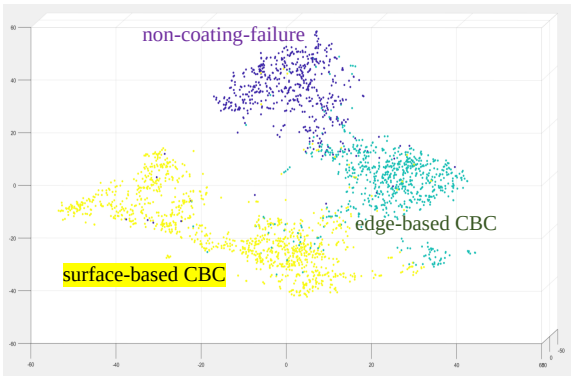


Figure 4: Feature visualization in t-SNE

Figure 5 represents a high-level features learnt from the three categories of extracted CBC features using deep dream. It shows that features for CBC on surfaces are some red dots with sharp gradient decent, while features for CBC on edges/welds are red dots along lines which means the corrosion happens on edges/welds. The non CBC features do not show any red corrosion color or texture. After extracting the R-CNN features, softmax [13] is used for multi-class feature classification and prediction.

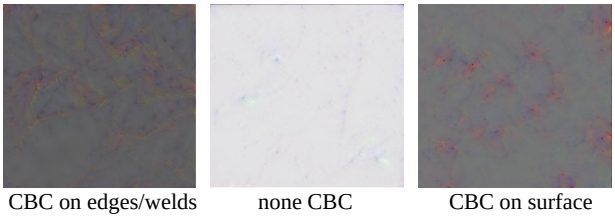


Figure 5: Feature visualization through deep dream

3.2 Feature Learning and Prediction

In the study, 1900 images with CBC were labeled for feature extraction; 12,184 features were extracted and divided into three broad categories, namely edge CBC, non-CBC and surface CBC. A random selection of 2437 features (20% of total features) was used for verification. In the experiment, vgg19 was chosen for network training. Stochastic gradient descent with momentum was used for loss function optimization; the learning rate was set to 1e-6; the maximum number of epochs for training was 20; and a mini-batch with 487 observations was used at each iteration. The max iteration was 9740. Validation frequency was set to 487 iterations duration. Validation patience value, the number of times that the loss on validation set can be larger than or equal to the previously smallest loss before network training stops, was set to five. During the experiment, the training process stopped when it reached the final iteration. The confusion matrix in Figure 7 demonstrates the verification accuracy. The total recognition accuracy rate observed was 89.54%.

The recognition accuracy can be improved by increasing the dataset. Continuous data augmentation will benefit the development of the CBC inspection system, improve CBC generalization and increase accuracy.

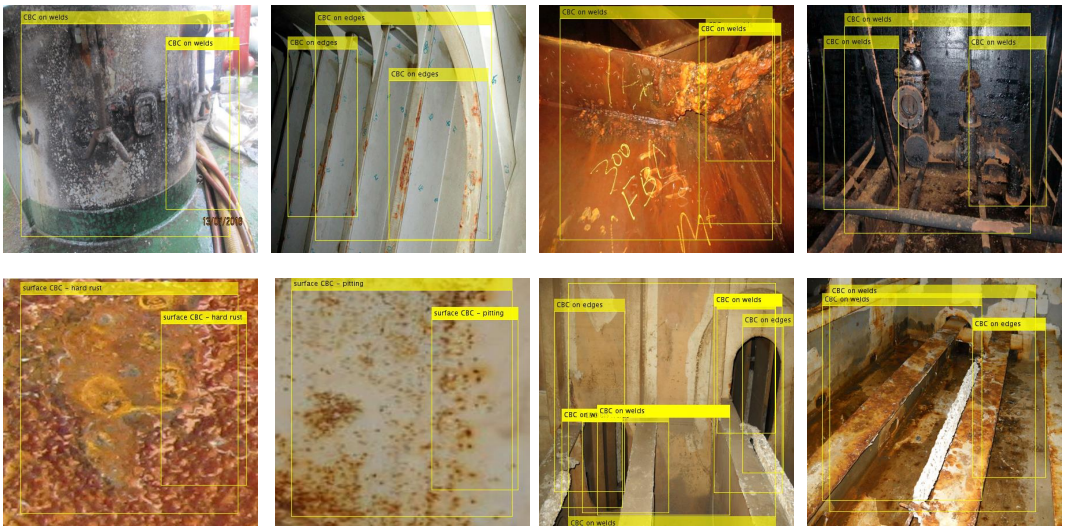


Figure 6: Detected CBC

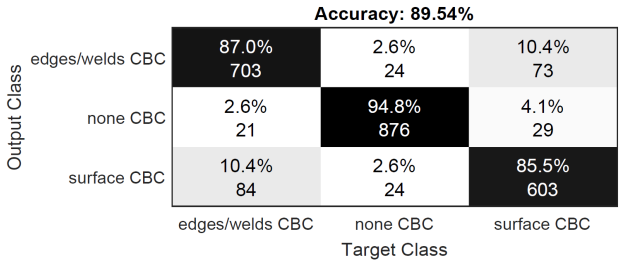


Figure 7: Feature prediction result

3.3 Feature Detection

Randomly generated bounding boxes are used to propose ROIs for different categories of CBC prediction. The predicted CBC ROIs are reconstructed with background removal. Instance-aware segmentation was developed for CBC instances' background removal.

The predicted CBC ROIs are reconstructed with background removal. Then, the reconstructed image is sent to the next step for active segmentation [8] and CBC measurement. A database with thousands of photos related to coating failure and corrosion was built up, which is expanding continuously as additional photos are collected from inspections. The system accurately identifies the main types of features, including surface CBC and edge CBC. Figure 6 shows classic examples of ROIs for CBC evaluation to speed up the inspection process. The CBC measurement is more precise and consistent than the ground truth images indicated in [9]. The ground truth images' assessment results are visually estimated by human surveyors and there is bias between different surveyors.

3.4 CBC Measurement

For CBC measurements, hue, saturation and value (HSV) color space is chosen to represent CBC color features. The extracted HSV data are used to identify possible CBC pixels to generate seeds for CBC detection. The pre-fetched HSV data contain most of the necessary information for pixel-wise CBC measurement. Our system gives users the flexibility to choose how they want to obtain the image that requires analysis. They can choose to either "Upload" or "Capture" an image with the device's camera. The algorithm 1 describes the detailed work flow. The system accurately identified the main types of CBC characteristics, including surface CBC and edge CBC. The study provides a comprehensive A-CAS system for marine and offshore industries. After post-analysis of the RGB image data (448 * 448 bytes), the system was able to promptly generate the inspection report containing field measurement data and the analysis result, as shown in Figure 8. A web-server for data communication will be developed to integrate the local server and client in autonomous on-line process. With the integration of all the techniques developed in this study, the AI-facilitated drone inspection system is able to assist surveyors to increase their efficiency during vessel condition inspection.

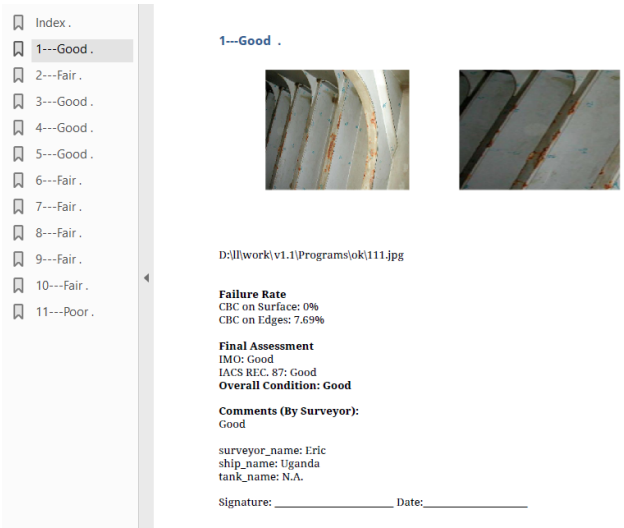


Figure 8: Report generation

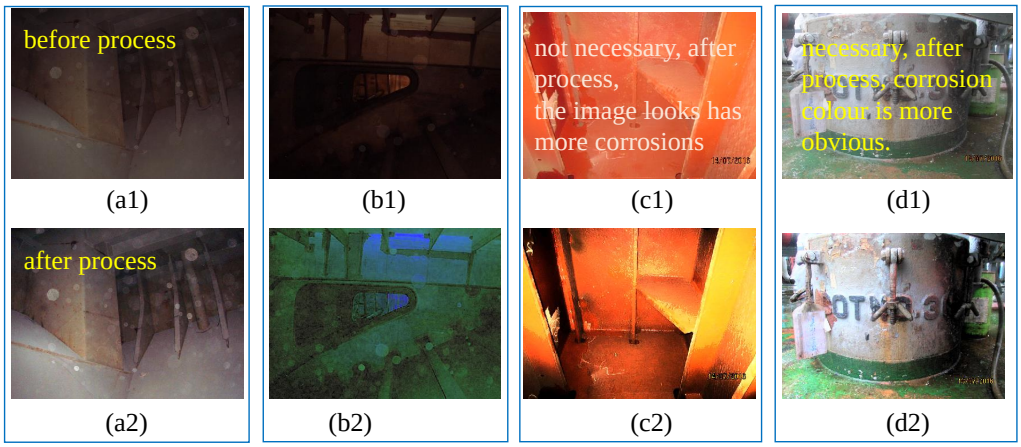


Figure 9: Image Enhancement

Algorithm 1: CBC Assessment System

Input: The original image
Output: CBC assessment result

Input an image and convolve network through different CNN layers;
Propose relative ROIs;
ROIs pooling;
Prediction through fully connected layer with softmax;
Morphological methods for edge CBC analysis, output processed image;
If an edge is discovered
    Edge CBC ROIs reconstruction;
    Hough transform finds lines;
    Active selection for propagation;
    Edge CBC quantitative measurement;
  Endif;
If surface CBC is detected
    Surface CBC ROIs reconstruction;
    Active semantic segmentation;
    Surface CBC quantitative measurement;
  End-if;
Adjust the weight for CBC on welds, if any;
Overall condition grading;
Generate report;
The end

The application program can be installed and operated on portable tablets and laptops. The CBC assessment report is automatically generated according to international standards such as IACS Recommendation 87 and IMO recommendations [14] (such as IMO MSC.1/Circ.1330 and IMO MSC.1/Circ.1399). Our program's results are consistent with the ground truth measurements of the surveyors. Our system continues to collect field data to optimize our algorithms and correlate the results with the ground truth results obtained by traditional methods.

4 Active Human Intervention to Improve Accuracy

The section above introduced a fully automated CBC assessment system (A-CAS) for effective coating failure inspection. This method is more suitable for inspecting large areas by capturing and analyzing images of the target area than the surveyor's existing manual inspection solution. However, the machine generated result is not 100 percent correct, and active human intervention and identification can correct and re-adjust these imperfect results. The following section describes how human intervention can fine-tune the result with the help of relevant advanced image processing methods.

4.1 Image Enhancement

Under normal circumstances, lighting will not affect the CBC assessment results. However, non-ideal lighting conditions, such as too high or too low light intensity, can affect the CBC assessment. So a function for adjusting brightness is implemented to facilitate CBC inspection. As shown in Figure 9 (a1) and (a2), we noticed that this function is helpful in certain cases.

Another affiliate function is Gamma correction. By exploiting the non-linear approach of human perception of light and color, Gamma correction works by optimizing the use of bits when encoding images, or transferring the bandwidth of an image. Figure 9 (b1) and (b2) show an example of before and after the Gamma correction process. It shows that this feature is also useful.

The third image enhancement function we implemented is histogram equalization. Histogram equalization was developed to increase contrast. The effect of histogram equalization is demonstrated in the two sets of examples in Figure 9 (c1) and (c2), (d1) and (d2). For CBC, whether histogram equalization helps to improve accuracy is situational dependent. For example, Figure 9 (c1) and (c2) shows that histogram equalization is not necessary, as the dust in Figure 9 (c1) is converted

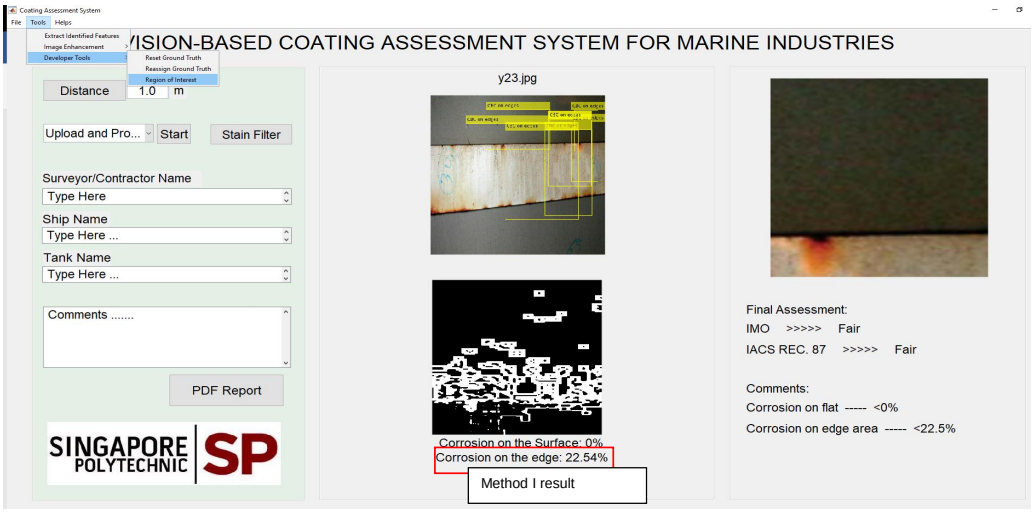


Figure 10: Method I: Auto-report generation

to the CBC color dataset, and the image looks more corroded after the histogram is equalized. For Figure 9 (d1) and (d2), the image before processing, Figure 9 (d1) has low brightness contrast, while Figure 9 (d2) shows that the brightness contrast has increased after processing, and CBC colors in the image better correspond to corrosion colors in the CBC color dataset. Therefore, histogram equalization is necessary for case (d1).

4.2 Ground Truth Adjustment

For special CBC color dataset adjustment, the reassign/reset ground truth function was developed as shown in Figure 11 for result fine-tuning. New special type of HSV values for CBC color can be added to CBC color dataset for CBC measurements.

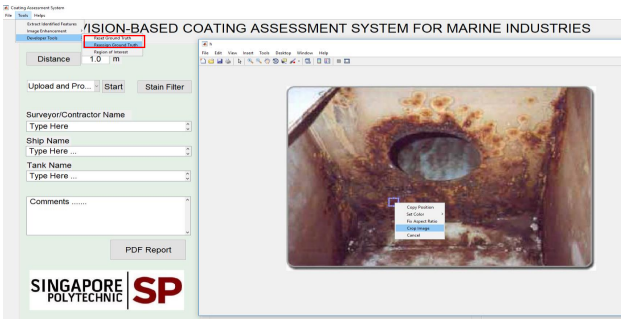


Figure 11: Ground truth adjustment

4.3 Active ROI Selection

There are two ways the system can measure edge CBC. Method I is the system default and the result is auto-generated by the system after input of images through the "Upload and Process", "Capture and Process" or "Batch Processing" buttons as shown in Figure 12. Figure 10 demonstrates a typical result based on method I. It shows the edge CBC is detected with 22.54%. The ROIs of CBC on edges which are auto-proposed by the CBC detection system is not perfect.

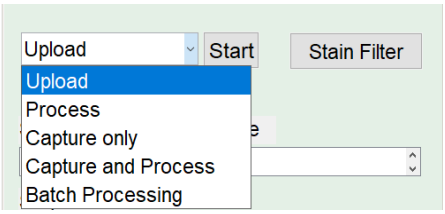


Figure 12: Method I: Auto report generation

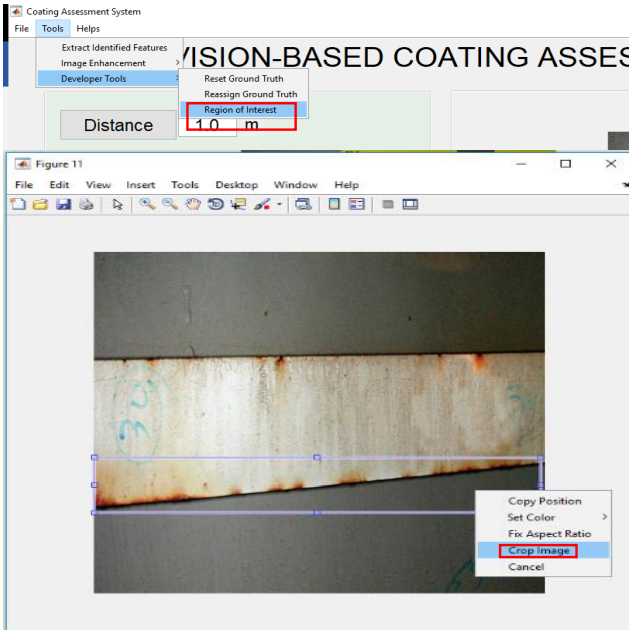


Figure 13: Method II: active ROI selection

Therefore, in order to align with human experience, active ROI selection method, i.e. method II, is developed to let machine interact with human and learn from human. It involves hard assignment of ROI especially for CBC on edges. As shown in Figure 13, the user needs to choose the "Tools" drop-down menu, followed by "Development Tools" and then click "Region of Interest" for ROI selection and edge CBC measure-

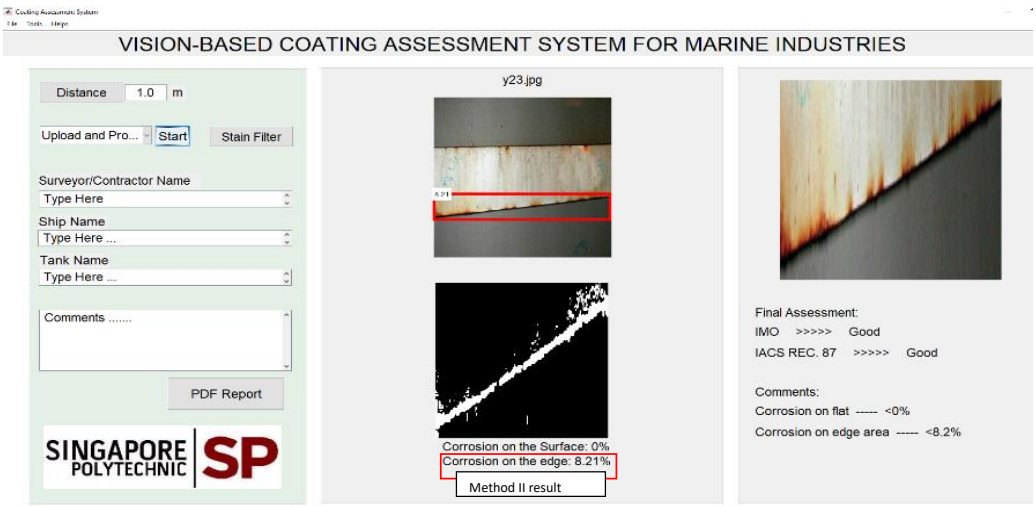


Figure 14: Method II result

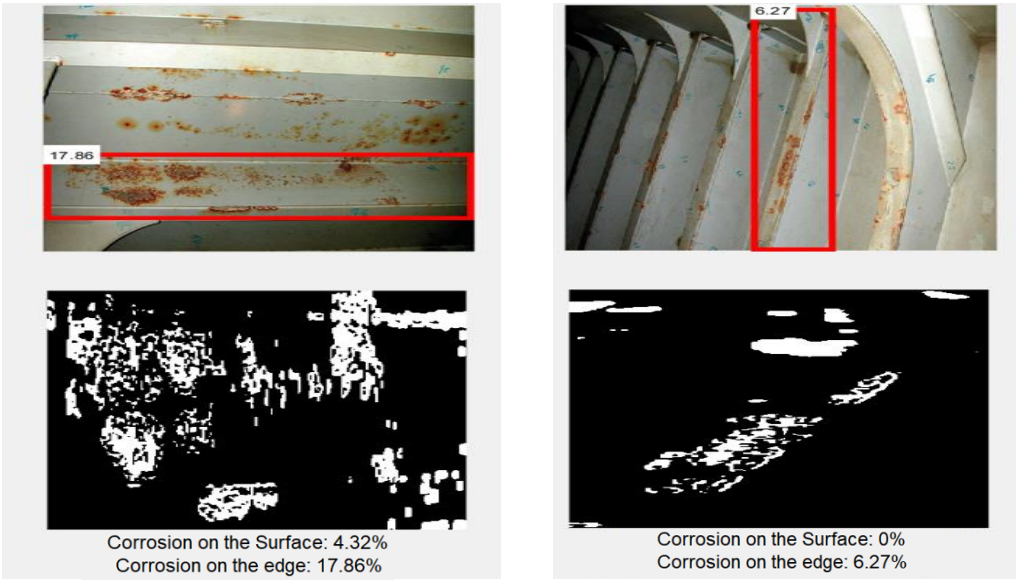


Figure 15: Typical results based on Method II

ment. After that, a new figure pops out to show the original image, where the user can draw a rectangle box on the image to select a ROI, and then right-click to "crop image". Afterwards, the edge CBC measurement result will be shown in the main GUI in Figure 14. This edge CBC result is 8.21%, which is more accurate than method I. Figure 15 displays two typical results by using method II. The results are more in line with human surveyors' experience.

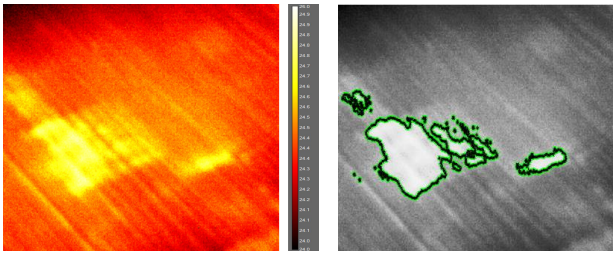
5 Non-destructive Hidden Corrosion Assessment

Hidden corrosion behind coated surfaces is a problem that could cause dangerous failure, causing shut down of production processes and high cost to the marine and offshore industry [15, 16]. At present, hidden corrosion can only be visually observed by human eyesight after cutting and/or coating removal. Compared to existing solutions, active infrared thermography (IRT) is relatively faster and non-destructive. Combined with algorithms for auto CBC detection, condition monitoring with active IRT can maintain higher precision and quality of assessment for consistent defect detection and measurement. While ultrasonic methods are limited to inspection of small areas, active IRT is fit for inspecting large areas, producing thermal images of the target. Active IRT can be used as an auxiliary tool for rough defect area positioning, followed by ultrasonic inspection for precise defect depth measurement. It can also be applied to risk management for cost savings and improving safety in the maritime industry. In the last 10 years, different research institutes have studied active IRT, but this technique has not been widely taken up by industry. The major problem is the long heating duration, making it unfit for on-site assessment. But with advances

in heating technology and sensitivity of thermal cameras, active IRT has potential to "see through" more advanced materials [17, 18, 19].

In this project, a mid-wave thermal imaging camera (FLIR a6702sc) and a halogen heater are chosen for instant heat generation to collect data in near real time for post image processing and analyzing [20]. By using a mid-wave thermal imaging camera FLIR A6703sc in combination with a 1500W halogen heater, the corrosion area could be imaged after only 5-10 seconds of heating. After removing the heat source, a temperature difference develops between the corroded and non-corroded areas, and the thermal contrasts as shown in Figure 16 (a) are captured by the thermal camera. In addition, the video captured is passed through machine learning algorithms for corrosion prediction and assessment. Figure 16 (b) is the processed result of hidden corrosion behind the coating for assessment.

Active infrared thermal imaging developed to detect hidden CBCs can be used to examine areas that are not identifiable by visual inspection. RBG imaged based CBC detection is used for 1st round screening test. Thermal camera is used for 2nd round hidden corrosion assessment once necessary. This will greatly increase the inspection effectiveness and help to predict and reduce the potential risk caused by the hidden corrosion.



(a) hidden coating corrosion (b) analyzed result

Figure 16: Hidden corrosion

6 Summary

The CNN-based automated CBC assessment system developed includes new coating breakdown and corrosion assessment algorithms with higher accuracy to facilitate quantitative data analysis for effective coating failure inspection. Compared to existing manual inspection solutions by surveyors, this method is more suitable for inspecting large areas by means of capturing and analyzing pictures/videos of the target areas.

The system accurately identified the main types of CBC characteristics, including surface CBC and edge CBC. The study provides a comprehensive A-CAS system for marine and offshore industries. A coating condition evaluation report is automatically generated according to international standards such as IACS Recommendation 87 and IMO recommendations. The detection system uses deep TLCAF technology to automate CBC assessment, and an instance-aware semantic segmentation method for CBC measurement and grading was developed. The three main types of CBCs can be distinguished and used for CBC prediction through the last fully connected layer. A model for coating failure prediction is generated, followed by CBC segmentation, CBC quantitative measurement and grading. The developed A-CAS system is going to improve maritime coating inspection works faster and smarter. With the assistance of IRT, the presence of corrosion behind protective coatings will be identifiable. Additionally, the non-destructive approach eliminates the need to remove the coating system.

The developed A-CAS system will make maritime coating inspection safer and more efficient. Shipping companies will be able to save on docking costs as well as eliminate losses incurred from goods left on the dock. The process of maritime coating inspection is set to have a makeover in the near future.

The days of cumbersome CBC inspections are gone forever as surveyors' jobs become highly automated. This automated corrosion assessment system can also be implemented in other industries that also face the problem of inefficient coating inspection. Such industries include the aviation, railway and building industries.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The author would like to thank Dr. Hai Gu of the American Bureau of Shipping (ABS) for his technical advice and data support. In addition, we are very grateful to Mr. Kelvin Lim Chee Quan, Mr. Ma Yiheng, Mr. Donvis Nguyen, Mr. Yeo Eng Hoe Jason and Mr. Duncan Goh Yitang for their contributions to the study. This work was funded by the Singapore Maritime Institute (grant no. SMI-2015-OF-05) and conducted at SP.

References

[1] L. Liu, E. Tan, Y. Zhen, X. J. Yin, Z. Q. Cai, Ai-facilitated coating corrosion assessment system for productivity enhancement, in: 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2018, pp. 606–610.

[2] M. Jahanshahi, S. Masri, Effect of color space, color channels, and sub-image block size on the performance of wavelet-based texture analysis algorithms: An application to corrosion detection on steel structures, in: Computing in Civil Engineering (2013), 2013, pp. 685–692.

[3] G. Ji, Y. Zhu, Y. Zhang, The corroded defect rating system of coating material based on computer vision, in: Transactions on Edutainment VIII, Springer, 2012, pp. 210–220.

[4] M. Siegel, P. Gunatilake, G. Podnar, Robotic assistants for aircraft inspectors, Industrial Robot: An International Journal 25 (6) (1998) 389–400.

[5] B. Zaidan, A. Zaidan, H. O. Alanazi, R. Alnaqeib, Towards corrosion detection system, International Journal of Computer Science Issues 7 (3) (2010) 33–36.

[6] A. Ortiz, F. Bonnin-Pascual, E. Garcia-Fidalgo, et al., Vision-based corrosion detection assisted by a micro-aerial vehicle in a vessel inspection application, Sensors 16 (12) (2016) 2118.

[7] L. Liu, R.-J. Yan, V. Maruvanchery, E. Kayacan, I.-M. Chen, L. K. Tiong, Transfer learning on convolutional activation feature as applied to a building quality assessment robot, International Journal of Advanced Robotic Systems 14 (3) (2017) 1729881417712620.

[8] S. Dutt Jain, K. Grauman, Active image segmentation propagation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2864–2873.

[9] A. American Buireau of Shipping, Guidance notes on the inspection maintenance and application of marine coating system third edition, Book, 2007.

[10] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[11] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[12] L. V. D. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[13] G. Hinton, R. Salakhutdinov, Replicated softmax: an undirected topic model, in: Advances in neural information processing systems, 2009, pp. 1607–1614.

[14] M. Resolution, 215 (82),"performance standard for protective coatings for dedicated seawater ballast tanks in all types of ships and double-side skin spaces of bulk carriers", IMO, London, UK.

[15] P. Traverso, E. Canepa, A review of studies on corrosion of metals and alloys in deep-sea environment, Ocean Engineering 87 (2014) 10–15.

[16] Z. Wang, Y. Cong, T. Zhang, et al., Effect of hydrostatic pressure on the pitting corrosion behavior of 316l stainless steel, Int J Electrochem Sci 9 (2014) 778–798.

[17] Y. Hung, Y. S. Chen, S. Ng, L. Liu, Y. Huang, B. Luk, R. Ip, C. Wu, P. Chung, Review and comparison of shearography and active thermography for nondestructive evaluation, Materials Science and Engineering: R: Reports 64 (5) (2009) 73–112.

[18] J. R. Brown, H. Hamilton, Quantitative infrared thermography inspection for frp applied to concrete using single pixel analysis, Construction and Building Materials 38 (2013) 1292–1302.

[19] L. Liu, I.-M. Chen, E. Kayacan, L. K. Tiong, V. Maruvanchery, Automated construction quality assessment: A review, in: Mechatronics and its Applications (ISMA), 2015 10th International Symposium on, IEEE, 2015, pp. 1–6.

[20] C. Duberstein, D. Virden, S. Matzner, J. Myers, V. Cullinan, A. Maxwell, Automated thermal image processing for detection and classification of birds and bats, Offshore Wind Technology Assessment.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 488-491 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Difference in Speech Analysis Results by Coding

Yasuhiro Omiya[*,1], Naoki Hagiwara[2], Takeshi Takano[2], Shuji Shinohara[3], Mitsuteru Nakamura[4], Masakazu Higuchi[4], Shunji Mitsuyoshi[3], Hiroyuki Toda[5], Shinichi Tokuno[4]

[1]*PST Inc., Industry & Trade Center Building 905, 2 Yamashita-cho, Naka-ku, Yokohama, Kanagawa 231-0023 Japan*

2Research and Product Development, PST Inc., 231-0023, Japan

[3]*Mathematical Engineering of Morality Emotions Graduate School of Engineering, The University of Tokyo, 113-0033, Japan*

[4]*Verbal Analysis of Pathophysiology Graduate School of Medicine, The University of Tokyo, 113-0033, Japan*

[5]*Department of Psychiatry, National Defense Medical College, 359-8513, Japan*

ARTICLE INFO

Article history:
Received: 07 August, 2018
Accepted: 08 October, 2018
Online: 25 October, 2018

Keywords:
Vocal analysis
Voice
Mental health care
Coding impact

ABSTRACT

Mental health disorder is becoming a social problem, and there is a need for technology that can easily check for states of stress and depression as a countermeasure. Conventional methods of diagnostic support and screening include self-administered psychological tests and use of biomarkers. However, there are problems such as burden on subjects, examination costs, dedicated reagents and equipment required for examinations, and reporting bias. On the other hand, voice-based evaluations are advantageous in terms of providing diagnostic support for physicians. They are non-invasive, do not require special and exclusive equipment, and can therefore be easily conducted remotely. We are pursuing the research and development of the Mind Monitoring System (MIMOSYS), which estimates the state of mental health from voice. Recorded audio is often compressed for efficient storage and transmission. However, there are concerns regarding the effects of deterioration of sound quality on analysis by MIMOSYS. Therefore, this study aims to verify the influence of the deterioration of voice quality due to coding on MIMOSYS analysis. As a verification experiment, coding was applied on the recording of 704 subjects reading 17 fixed phrases, assuming compression for transmission and storage. Then, the results of MIMOSYS analysis before and after encoding were compared. A strong correlation was observed before and after encoding, suggesting that MIMOSYS analysis is also valid for voice to which coding was applied.

1. Introduction

This paper is an extended paper from IEEE ICIIBMS 2017 (The International Conference on Intelligent Informatics and Biomedical Sciences) [1].

Mental health disorder is becoming a social problem, and there is a need for technology that can easily check for states of stress and depression as a countermeasure.

Conventional methods of diagnostic support and screening include self-administered psychological tests and use of biomarkers. However, there are problems such as burden on subjects, examination costs, dedicated reagents and equipment required for examinations, and reporting bias.

On the other hand, voice-based evaluations are advantageous in terms of providing diagnostic support for physicians. They are non-invasive, do not require special and exclusive equipment, and can therefore be easily conducted remotely.

We are pursuing the research and development of the Mind Monitoring System (MIMOSYS), which estimates the state of mental health from voice [2]. Recorded audio is often compressed for efficient storage and transmission. However, there are concerns regarding the effects of deterioration of sound quality on analysis by MIMOSYS. Therefore, this study aims to verify the influence

*Yasuhiro Omiya, , +81-45-263-9346 & omiya@medical-pst.com

of the deterioration of voice quality due to coding on MIMOSYS analysis.

2. Materials and Methods

2.1. Vocal Analysis by MIMOSYS

Generally, a healthy person expresses emotions such as joy, anger, sadness, or pleasure depending on changes in their surroundings. On the other hand, when stress accumulates and the person becomes depressed, the expression of emotions declines. MIMOSYS is based on the sensibility technology ST (AGI Inc.) [3] which measures emotion from voice. Short-term mental health index ("vitality") and long-term mental health index ("mental activity") are analyzed from the balance of five combined emotional indices, which are the four emotions "calmness", "anger", "joy", and "sorrow" obtained by voice analysis, and "excitement" represented by the strength of brain emotional activity. ST software detects voice activity based on volume as shown in Figure 1. For each utterance, feature values such as pitch, power, and deviation in power are calculated by a robust fundamental frequency and intonation detection method. Based on these feature values, they are classified into emotional elements of "calmness", "anger", "joy", and "sorrow" through decision tree logic. The degree of "excitement" is also analyzed as the strength of brain emotional activity. Each emotion detected from the input voice is an integer value from 0 to 10, and the strength of "excitement" is output as an integer value from 1 to 10.

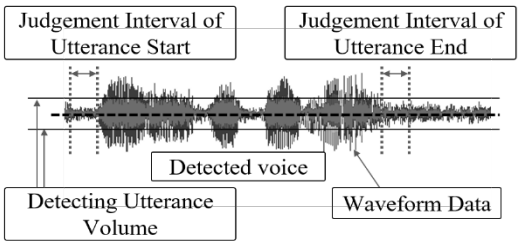


Figure 1. Voice activity detection

MIMOSYS calculates "vivacity" and "relaxation" as intermediate feature values, in the course of its analysis. Vivacity is calculated from "joy" and "sorrow", and relaxation is calculated from "calmness" and "excitement". The short-term mental health index "vitality" is then calculated from these, and the mid-long term mental health index "mental activity" is calculated from the "vitality" accumulated over two weeks. The flow of MIMOSYS processing is shown in Figure 2.

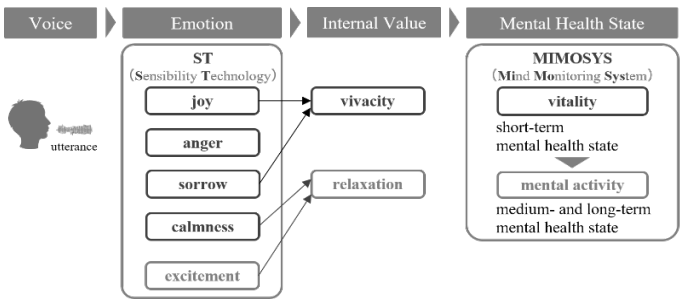


Figure 2. The flow of MIMOSYS processing

2.2. Beck Depression Inventory

The Beck Depression Inventory (BDI) test is widely used for depression screening and consists of 21 questions on how the subject has been feeling in recent days, including on the day. Each question has four options, and the total score determines the severity of depression. Table 1 shows cutoff values according to reference [4]. In this study, since only healthy subjects were targeted, only the voices of subjects with a score of 9 or less were used.

Table 1. cutoff scores for BDI.

Total BDI score	*Level of depression*
0 – 9	indicates minimal depression
10 – 18	indicates mild depression
19 – 29	indicates moderate depression
30 – 63	indicates severe depression

2.3. Voice Collection

We prepared voice (96 kHz, 24 bit, wav file) collection which collected using the Portable Recorder R-26 (Roland, Japan) and a pin microphone ME52W (OLYMPUS, Japan) from male and female subjects (n=986) in the assembled soundproof shelter S-909 (Starlight, Japan). The voices were collected when subjects reading 17 fixed phrases. Additionally, BDI test was conducted before the experiment. Then we excluded abnormal participants using BDI score. There were 704 subjects with screened data used in the analysis. Then we down-converted to 11,025 Hz, 16 bits, which is the input format of MIMOSYS and is used as original voice in this research. Hereafter, audio processing was done using "SoundExchange" [5] and "FFmpeg" [6] software with default coding settings.

This study was approved by the Research Ethics Review Committee. Informed consent was obtained from the subjects.

2.4. Experiment

We prepared data collection of recording format corresponded to linear PCM with a sampling frequency of 11,025 Hz, and a quantized bit number of 16 bits as "Base voice(11k)".

Next, "Base voice(11k)" was converted to a sampling frequency of 8,000 Hz and a quantized bit number of 16 bits simulating transmission across a telephone line. Then, coding methods "G.723.1(MP-MLQ)", "G.723.1(ACELP)", and "G.729", which are based on ITU-T recommendations [7,8], were applied to "Base voice(8k)". Additionally, coding methods "AAC", "MP3", and "WMA", were applied to "Base voice", simulating recording and/or archiving. "Base voice" and each coded voice were decoded and assigned an analyzed "vitality" score through MIMOSYS.

Finally, the "vitality" score before and after coding using each coding method was compared by correlation evaluation. We also compared the analyzed "vitality" score of "Base(11k)" and "Base(8k)" in order to confirm the influence of resampling (down-

sampling and up-sampling). We analyzed the data using Microsoft Excel 365 (Microsoft Corp.).

3. Results and Discussion

Figure 3-5 is a scatter diagram of "vitality" before and after encoding when "G.723.1(MP-MLQ)", "G.723.1(ACELP)", and "G.729" is used as the voice coding method, assuming transmission over a telephone line. Figure 6-8 is a scatter diagram of "vitality" before and after encoding when "AAC", "MP3", and "WMA" is used, assuming recording and archiving. Figure 9 is a scatter diagram of "vitality" before and after resampling. In addition, Table 2 shows correlation coefficients of "vitality" before and after encoding.

Table 2. correlation coefficients of "vitality" before and after encoding.

Compared	*Coding method*	*Coefficient of correlation*
Base(8k)	"G.723.1(MP-MLQ)"	0.882
Base(8k)	"G.723.1(ACELP)"	0.740
Base(8k)	"G.729"	0.802
Base(11k)	"AAC"	0.959
Base(11k)	"MP3"	0.972
Base(11k)	"WMA"	0.962
Base(11k)	Base(8k)	0.780

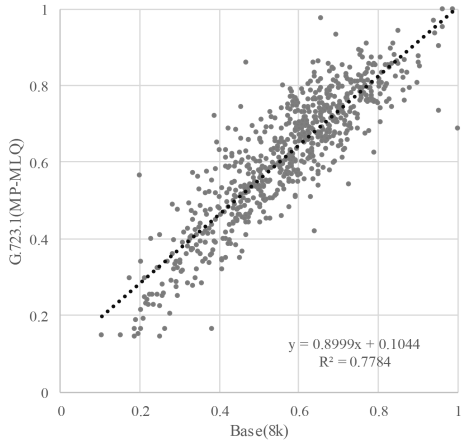


Figure 3. Scatter diagram of "vitality" encoding with "G.723.1(MP-MLQ)".

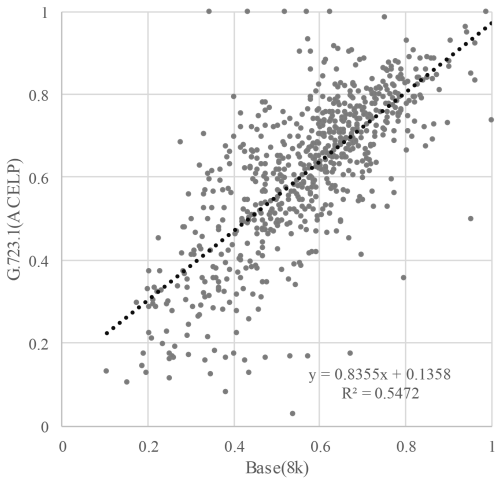


Figure 4. Scatter diagram of "vitality" encoding with "G.723.1(ACELP)".

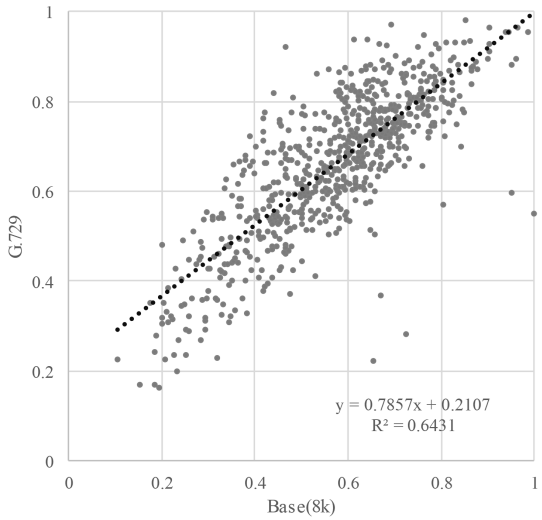


Figure 5. Scatter diagram of "vitality" encoding with "G.729".

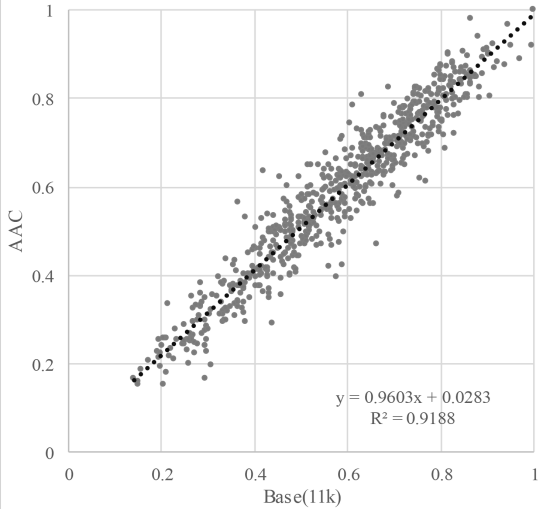


Figure 6. Scatter diagram of "vitality" encoding with "AAC".

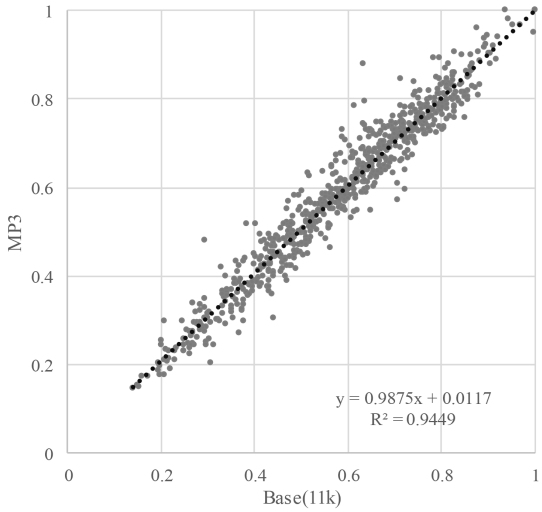


Figure 7. Scatter diagram of "vitality" encoding with "MP3".

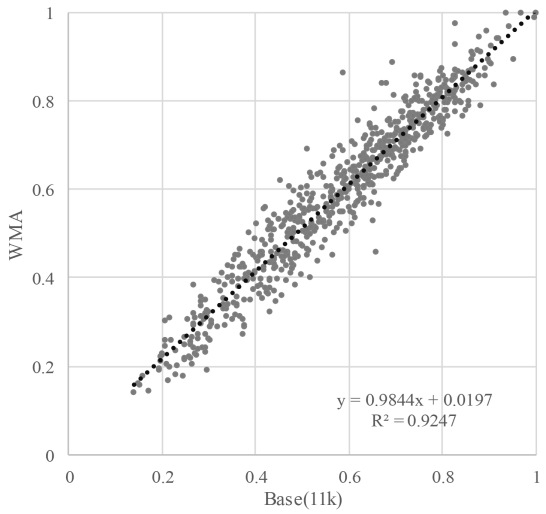


Figure 8. Scatter diagram of "vitality" encoding with "WMA".

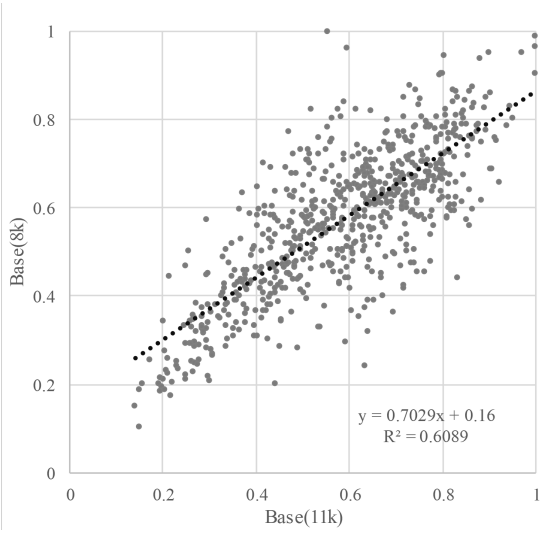


Figure 9. Scatter diagram of "vitality" before and after resampling.

Table 3. Average value of difference in "vitality" before and after coding.

Coding method	*Ave.± S.D.*	*Base(11k)*	*Base(8k)*
"G.723.1(MP-MLQ)"	0.619±0.166	0.033	0.047
"G.723.1(ACELP)"	0.613±0.184	0.028	0.042
"G.729"	0.660±0.160	0.074	0.088
"AAC"	0.590±0.181	0.005	0.019
"MP3"	0.589±0.184	0.004	0.018
"WMA"	0.596±0.185	0.011	0.024
Base(8k)	0.571±0.163	0.014	-

The coefficients of correlation before and after coding is extremely high at about 0.74. Coding methods of recording and/or archiving such as "AAC", "MP3", and "WMA" are extremely high at about 0.95. Further, as shown in Table 3, the results suggested that the influence of sampling frequency conversion on "vitality" tends to be higher after coding. This result show that the influence on MIMOSYS analysis of sound degradation, either due to coding for transmission across a telephone line or coding for recording or archiving, is minor and is therefore a useful method. In Figure 4, there are multiple anomaly points where "vitality" after coding by "G.723.1(ACELP)" is observed around 1 and less than 0.2. This is a peculiar feature of this coding method and use of other coding methods might be preferred. On the other hand, the coefficients of correlation before and after a sampling frequency conversion without coding causes deterioration. This deterioration is inevitable because the input format of MIMOSYS is 11,025 Hz while the input format of the general telephone line is 8,000 Hz.

Prior to this research, we conducted an examination of simulating transmission across a telephone line using coding methods of "u-law", "A-law", "G.726", and "G.723.1" [9]. In this study and previous research, results show other coding methods such as compression methods are useful after coding.

4. Conclusion

In this study, in order to verify the influence of deterioration of voice quality on the estimation of health condition using voice by MIMOSYS, the results of MIMOSYS analysis before and after applying coding to voice were compared and verified. The correlation coefficients of mental activity before and after the application of voice coding were high, and it was suggested that even voice compression coded for transmission via a telephone line or storage could be used. Furthermore, it was also suggested that the influence of sampling frequency conversion on vitality tends to be larger than applying coding.

In the future, we would like to further improve accuracy by considering sampling frequency conversion methods and value correction methods with less influence on MIMOSYS. We also want to improve the accuracy of the MIMOSYS algorithm itself.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

[1] Yasuhiro Omiya, Naoki Hagiwara, Takeshi Takano, Shuji Shinohara, Mitsuteru Nakamura, Masakazu Higuchi, Shunji Mitsuyoshi, and Shinichi Tokuno, "Difference in Speech Analysis Results by Compression," in International Conference on Intelligent Informatics and BioMedical Sciences (ICIIBMS) 2017, Okinawa, Japan, 2017.

[2] Shinichi Tokuno, "Pathophysiological Voice Analysis for Diagnosis and Monitoring of Depression," Understanding Depression, Springer, Singapore, 83-95, 2018.

[3] Shunji Mitsuyoshi, Fuji Ren, Yasuto Tanaka, and Shingo Kuroiwa, "Non-verbal voice emotion analysis system," Int. J. Innov. Comput. I., 2(4), 819–830, 2006.

[4] Beck, AT, Steer RA, Garbin MG J.: Psychometric properties of the Beck Depression Inventory Twenty-five years of evaluation. Clinical Psychology Review 8(1), 77–100, 1988.

[5] Chris Bagwell, SoX – Sound eXchange, Available at: <http://www.soundexchange.com/> (Accessed 12 Jul. 2018)

[6] FFmpeg Developers, "ffmpeg tool (Version 20170503-a75ef15) [software]," Available from: <https://www.ffmpeg.org/> (Accessed 28 Sep. 2018)

[7] ITU-T Recommendation G.723.1 (05/2006), Available at: < https://www.itu.int/rec/T-REC-G.723.1-200605-I/en> (Accessed 28 Sep. 2018)

[8] ITU-T Recommendation "G.729" (06/2012), Available at: <https://www.itu.int/rec/T-REC-"G.729"-201206-I/en> (Accessed 28 Sep. 2018)

[9] Naoki Hagiwara, Yasuhiro Omiya, Shuji Shinohara, Mitsuteru Nakamura, Masakazu Higuchi, Shunji Mitsuyoshi, and Shinichi Tokuno, "Difference in Voice Analysis Result by Pre- and Post- Processing of Telephone Line," IEEE EMBC 2017, Jeju, Korea, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 496-500 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Omni-directional Dual-Band Patch Antenna for the LMDS and WiGig Wireless Applications

Mourad S. Ibrahim[1,2,*]

[1]*Department of Communications and Networks Engineering, College of Engineering, Prince Sultan University, Riyadh, 11586, KSA*

[2]*College of Engineering, Modern Sciences and Arts University, 6th October City. Egypt.*

ARTICLE INFO

Article history:
Received: 20 September, 2018
Accepted: 16 November, 2018
Online: 23 December, 2018

Keywords:
Dual band
Fifth generation
LMDS
Omnidirectional pattern
WiGig

ABSTRACT

In this paper an omnidirectional dual band monopole antenna at 28 GHz and 60 GHz which is fit for indoor and outdoor wireless applications is developed. The proposed antenna consists of two rectangular patches with a T folded patch. The design, analysis, and optimization processes through this article are executed by the numerical method, Finite Element Method (FEM) and verified with another numerical method, Finite integration Technique (FIT). Good agreement between the results by these two simulators is obtained. The proposed antenna has achieved dual bands with omnidirectional patterns. The first band at 28 GHz is extended from 27.5 GHz to 28.958 GHz with 5.1 % bandwidth and total efficiency of more than 93% along the entire band which serves the LMDS band. The second band at 60 GHz is extended from 45.2 GHz to 84.4 GHz which serves the WiGig band with bandwidth of 60.6% and total efficiency of 85.5% along the entire band. The proposed antenna performance makes it a good candidate for the fifth generation (5G) applications.

1. Introduction

This paper is an extension of work originally presented in ACES [1]. More analysis and parametric studies have been done. The average data per person uses in telecommunication is rapidly increases over the past three decades [2]. This increase is more noticeable mostly with the outgrowth of the wireless communications. The main requirements for wireless communications are the ability to develop a low-cost, light-weight, and low-profile antennas to maintain good performance along a large bandwidth [2].

The shortage in the available global bandwidth has stimulated the reconnaissance of the under-utilized millimeter wave frequency band for the future wireless communications [3].

Multiband antennas have massive applications in millimeter band applications. A various techniques in the literature have been developed for multiband microstrip antennas as in [4–12]. For instance, a multilayer GaAs is described in [4] to achieve a multiband antenna operating at 35 GHz. In [5], two bands with less than 1.2% bandwidth with gains of -9 dBi and 1 dBi at 24 GHz band and 60 GHz band respectively is presented. A dual band antenna at 41 GHz / 52.2 GHz using a meta-resonator with pair of split ring resonators is introduced in [6] with bandwidth of 2 %, gain of 3.76 dBi, and efficiency of 71%. In [7], two different modes are obtained to get a dual band at 58 GHz and 77 GHz with gains of -2 dBi and 0.3 dBi, respectively. The achieved bandwidths at both bands are nearly 6%. A dual band centered at 24.5 and 35 GHz has been presented in [8] with only 1% bandwidth and less than 2.8 dBi gain using liquid crystal polymer. A coplanar hybrid dual band antenna at 83 GHz / 94 GHz with a slot in feeding line has been developed in [9]. In [10] an antenna array with Electromagnetic Band Gap [EBG] structure was used to develop a dual band at 28 / 38 GHz with bandwidth of less than 5.8%. A three layers of substrates have been used in [11] to design a Fabry-Perot cavity antenna operates at 36 GHz with high gain. In [12], L-shaped slots have been used to obtain a dual band 28 / 38 GHz slotted patch antenna.

In this paper, the design, optimization, and simulation of a monopole planar antenna are introduced. The antenna is optimized to operate at Ka – band (28 GHz) for Local Multipoint Distribution Service (LMDS) which currently investigated for the fifth generation mobile cellular [13], and the V – band (60 GHz) for Wireless Gigabit Alliance (WiGig) applications [14].

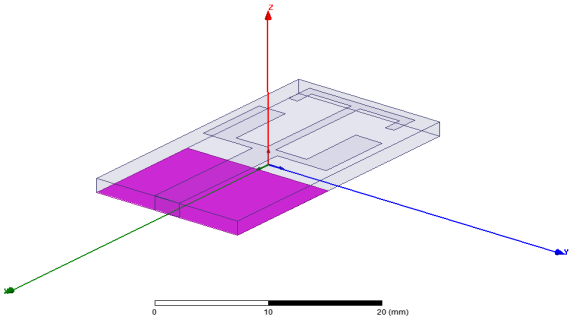
This paper is organized into four sections. Section 1 covers the introduction and literature review. Antenna structure and design is introduced in section 2. Parametric study of a dual band

*Mourad S. Ibrahim, Riyadh 11586, mrizk@psu.edu.sa

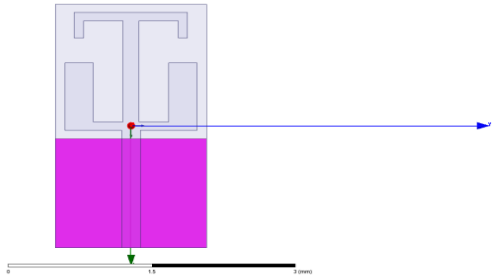
omnidirectional pattern circularly polarized wideband antenna is presented in section 3. The simulation results are investigated in Section 4. Finally, the conclusion is presented in section 5.

2. Antenna Geometry and Design

Figure 1 shows the antenna geometry in perspective and top view. The proposed antenna patch consists of a T-folded shape with two rectangular patches and partially grounded [15]. The antenna gives wide bandwidth with improved antenna performance. The patch mounted on substrate FR-4 with a relative permittivity of 4.4. Figure 2 illustrates the whole antenna dimensions. The substrate dimension is W_s by L_s and the partial ground is W_s by L_g. The dimensions of T-folded are L_{top}, L_{fold} , L_t W_t for top length, folded side length, mid length, and mid width. The two patches have dimensions of W_p by L_p.



(a)



(b)

Figure1. Antenna geometry a) Perspective view b) Top view.

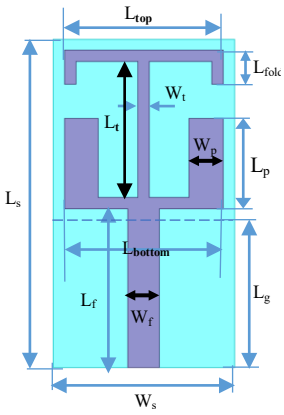


Figure 2. Antenna structure and dimensions.

The antenna feeding transmission line (TL) dimension is W_f by L_f. The TL characteristic impedance can be calculated by the following formulas [16]:

When $\frac{w}{h} \le 1$, the characteristic impedance is

$$Z_c = \frac{60}{\sqrt{\varepsilon_{reff}}} \ln\left[\frac{8h}{w} + \frac{w}{4h}\right] \quad (1)$$

where

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} \times \left\{\left[1 + 12\frac{h}{w}\right]^{-1/2} + 0.04\left[1 - \frac{w}{h}\right]^2\right\} \quad (2)$$

while when $\frac{w}{h} > 1$

$$Z_c = \frac{\frac{120\pi}{\sqrt{\varepsilon_{reff}}}}{\frac{w}{h} + 1.393 + 0.667 \ln\left[\frac{w}{h} + 1.444\right]} \quad (3)$$

where

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 12\frac{h}{w}\right]^{-1/2} \quad (4)$$

In the formulas, h represents the substrate height, w represents the TL width, and ε_{reff} represents the effective relative permittivity.

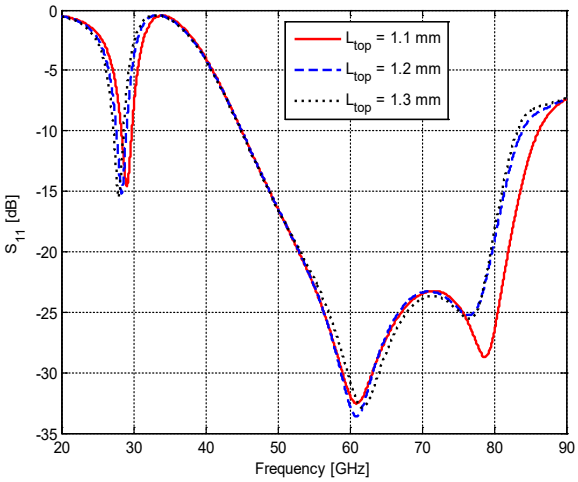


Figure. 3. The effect of changing L_{top} on S_{11}.

3. Parametric Study

The effect of varying antenna dimensions L_{fold}, L_{top}, L_{bottom}, L_P, W_P, and L_t using HFSS are shown in figures 3 to 8. The dimensions of the substrate, ground, and feeder are kept unchanged. With increasing L_{top} of the horizontal part of T-shaped, the lower resonant frequency is decreased with same bandwidth whereas the upper bandwidth is decreased as shown in Figure 3. Figure 4 illustrates the effect of varying L_{bottom} on the return loss. As noted from the figure, by increasing the L_{bottom}, the lower resonant frequency almost unaltered whereas the resonant frequency of the

upper band is increased with a little increase in the bandwidth. With increasing W_P, the upper and the lower resonant frequencies are almost unchanged as shown in Figure 5.

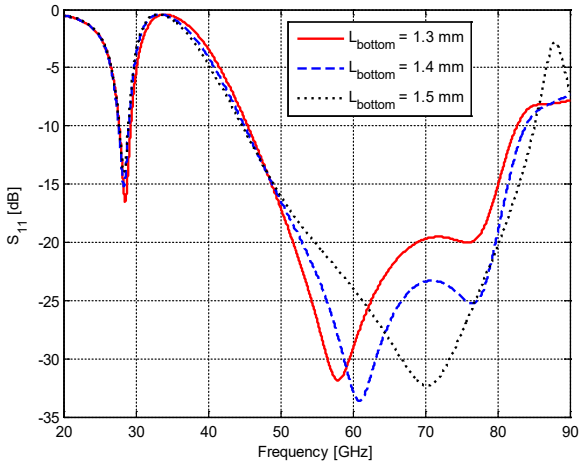


Figure. 4. The effect of changing L_{bottom} on S_{11}.

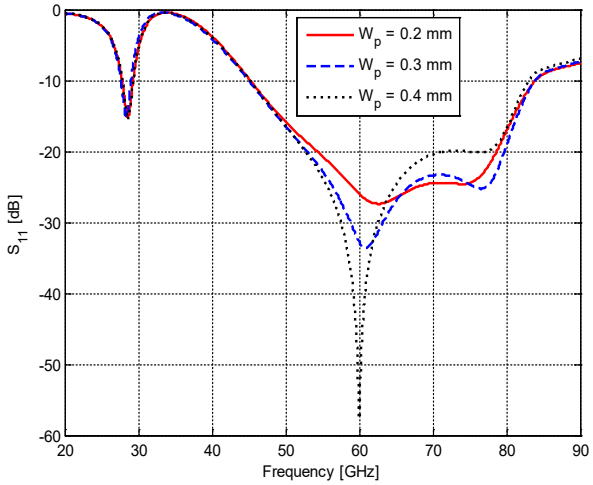


Figure. 5. The effect of changing W_p on S_{11}.

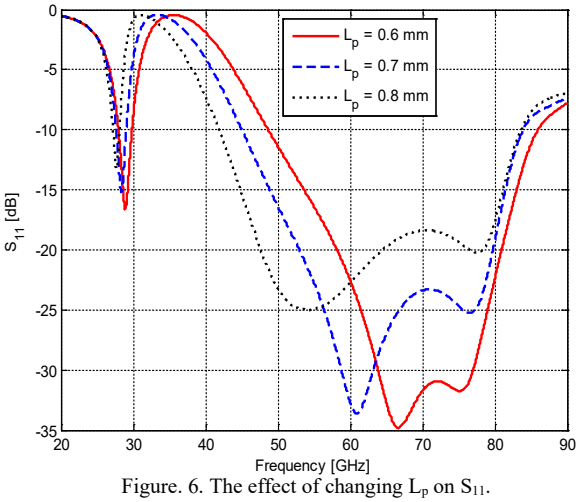


Figure. 6. The effect of changing L_p on S_{11}.

As the length of rectangular patch L_P increases, the upper and lower resonant frequency are decreased but the antenna matching becomes worse. Lower bandwidth is increases whereas the upper bandwidth is decreases as shown in Figure 6. Figure 7 illustrates the return loss varies with L_{fold}. As L_{fold} increases the lower resonant frequency decreases while upper resonant frequency unchanged whereas the lower bandwidth is not affected, the upper bandwidth is a little decreases as shown in Figure 7. With increasing L_t, the lower and the upper resonant frequencies are not affected whereas the upper bandwidth is decreases, the lower bandwidth is unchanged as shown in Figure 8.

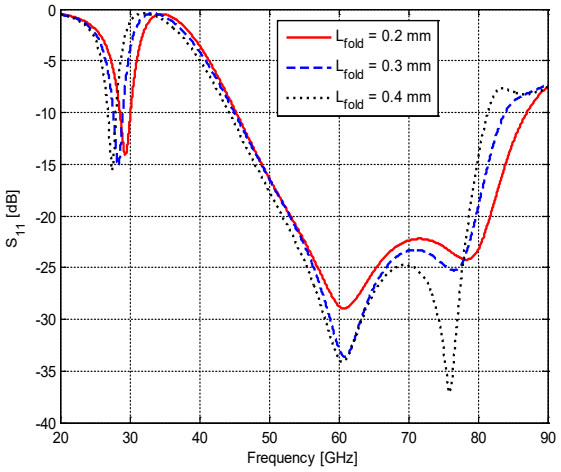


Figure. 7. The effect of changing L_{fold} on S_{11}.

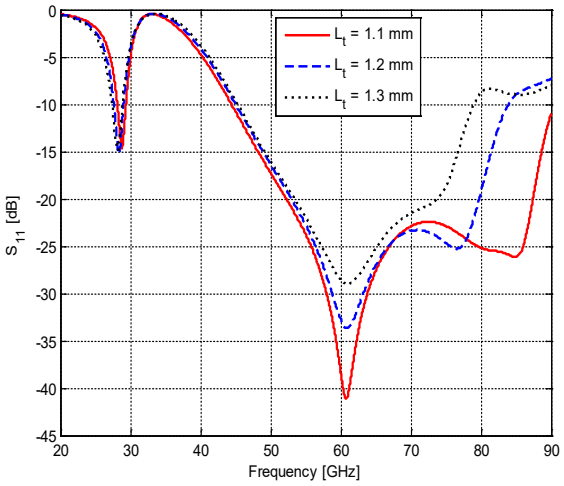


Figure. 8. The effect of changing L_t on S_{11}.

4. Antenna Simulation Results

The antenna dimensions are optimized to obtain an Omni-directional radiation patterns with acceptable performance at 28 GHz and 60 GHz. Table 1 contains the optimized dimensions for the proposed antenna in order to have a dual band.

The optimized antenna is simulated using two simulators based on FIT [17] and FEM [18] in order to validate the results. The S_{11} for the proposed antenna using FEM and FIT is shown in Figure 9. The FEM gives a better matching than FIT and the small deviation in the 60 GHz band is due to the different mesh sizes. As can be noted from Figure 9, the antenna well matched over the two bands and the impedance bandwidth for which $S_{11} \leq$ -10 dB in 28 GHz band is extended from 27.52 GHz to 28.96 GHz which

serves the LMDS band and in 60 GHz band is extended from 45.2 GHz to 84.4 GHz which serves the WiGig band.

Table 1: Dimensions of the proposed antenna in mm.

Parameter	Value	Parameter	Value
W_s	1.6	W_p	0.3
L_s	2.9	L_p	0.7
L_g	1.3	W_t	0.1
W_f	0.2	L_t	1.2
L_f	1.4	L_{top}	1.2
L_{fold}	0.3	L_{bottom}	1.4

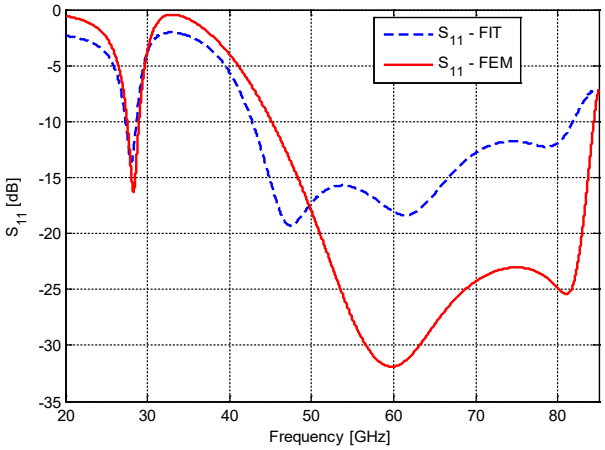


Figure. 9. Comparison of S_{11} for the optimized antenna by FIT and FEM.

The antenna resonates at 28.24 GHz with 1.44 GHz bandwidth (5.1 %) and at 64.76 GHz with wide bandwidth of 39.24 GHz (60.6%). The VSWR is shown in Figure 10 with VSWR ≤ 2 using FIT and FEM in the entire bandwidths of both bands.

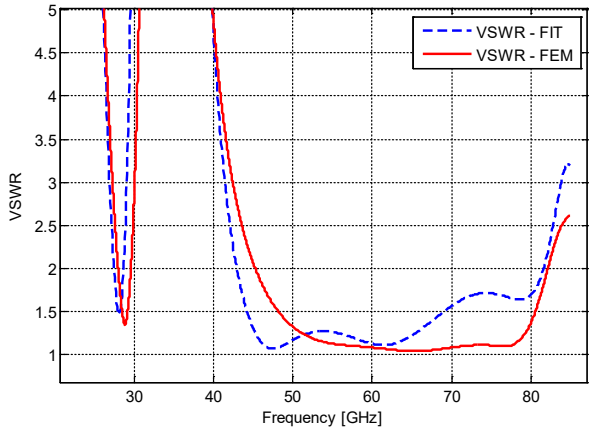


Figure. 10. VSWR using FIT and FEM.

The antenna real and imaginary input impedance for the proposed antenna is illustrated in Figure. 11. As can be noted from Figure 11, the antenna is well matched to 50 Ω TL since the real part of the input impedance is approximately 50 Ω while the imaginary part tends to zero along the entire bandwidths of both bands.

The radiation patterns at 28 GHz and 60 GHz are illustrated in Figure 12. The directivity pattern at 28 GHz with D_o of 2.28 dBi is illustrated in Figure 12(a) and at 60 GHz with D_o of 3.41 dBi is shown in Figure 12(b). The total antenna efficiency along the lower band is 93 % while along the upper band is 85.5 %.

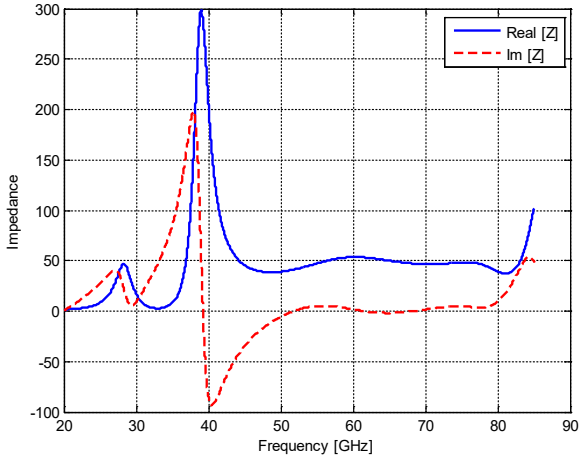
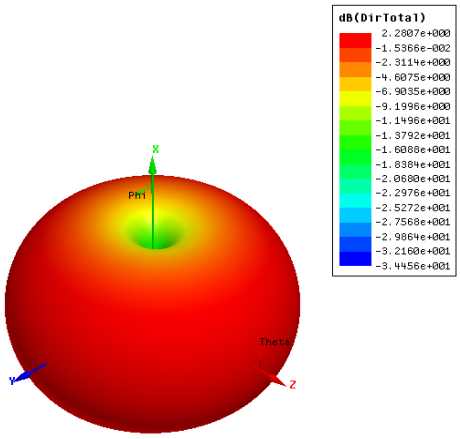
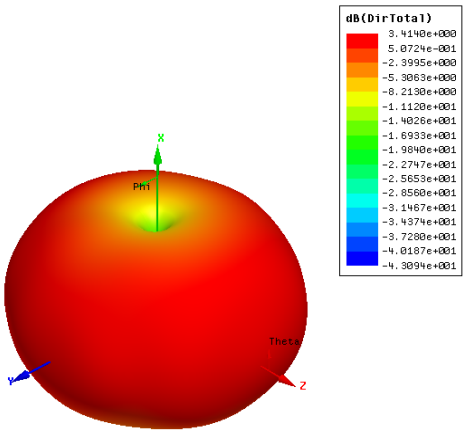


Figure. 11. The real and imaginary part of input impedance for the proposed antenna.



(a)



(b)

Figure 12. The directivity patterns at a) 28 GHz (D_o = 2.28 dBi) and at b) 60 GHz (D_o = 3.4 dBi).

5. Conclusion

In this paper, antenna with two rectangular patches and a T-shaped folded patch is designed, optimized, and simulated. The antenna exhibits two bands resonate at 28.24 GHz and 64.76 GHz with bandwidths of 1.44 GHz (5.1%), and 39.24 GHz (60.6%) respectively. The lower band (Ka-band) is suitable for LMDS while the upper band (V-band) with wideband is suitable for WiGig. The omnidirectional radiation pattern is obtained using partial ground plane with maximum directivities of 2.28 dBi and 3.414 dBi at the two bands respectively. The total efficiency along the entire lower and upper bands exceeds 85% which can be used in the fifth generation applications.

Acknowledgment

The author gratefully acknowledges the support of Prince Sultan University.

References

[1] M. S. Ibrahim, "Dual-band microstrip antenna for the fifth generation indoor/outdoor wireless applications," in *2018 International Applied Computational Electromagnetics Society Symposium (ACES)*, 2018, pp. 1-2.
[2] S. S. Jaco du Preez, *Millimeter-Wave Antennas: Configurations and Applications*: Springer International Publishing Switzerland, 2016.
[3] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access,* vol. 1, pp. 335-349, 2013.
[4] D. Sanchez-Hernandez, Q. H. Wang, A. A. Rezazadeh, and I. D. Robertson, "Millimeter-wave dual-band microstrip patch antennas using multilayer GaAs technology," *IEEE Transactions on Microwave Theory and Techniques,* vol. 44, pp. 1590-1593, 1996.
[5] H. Jie-Huang, W. Jin-Wei, C. Yi-Lin, and C. F. Jou, "A 24/60GHz dual-band millimeter-wave on-chip monopole antenna fabricated with a 0.13-μm CMOS technology," in *2009 IEEE International Workshop on Antenna Technology*, pp. 1-4, 2009.
[6] I. K. Kim and V. V. Varadan, "Electrically Small, Millimeter Wave Dual Band Meta-Resonator Antennas," *IEEE Transactions on Antennas and Propagation,* vol. 58, pp. 3458-3463, 2010.
[7] T. Y. Lin, T. Chiu, and D. C. Chang, "Design of Dual-Band Millimeter-Wave Antenna-in-Package Using Flip-Chip Assembly," *IEEE Transactions on Components, Packaging and Manufacturing Technology,* vol. 4, pp. 385-391, 2014.
[8] D. Lee and C. Nguyen, "A millimeter-wave dual-band dual-polarization antenna on liquid crystal polymer," in *2014 IEEE Antennas and Propagation Society International Symposium (APSURSI)*, pp. 775-776, 2014.
[9] S. Agarwal, N. P. Pathak, and D. Singh, "Concurrent 83GHz/94 GHz parasitically coupled defected microstrip feedline antenna for millimeter wave applications," in *2013 IEEE Applied Electromagnetics Conference (AEMC)*, pp. 1-2, 2013.
[10] N. Ashraf, O. Haraz, M. A. Ashraf, and S. Alshebeili, "28/38-GHz dual-band millimeter wave SIW array antenna with EBG structures for 5G applications," in *2015 International Conference on Information and Communication Technology Research (ICTRC)*, pp. 5-8, 2015.
[11] G. N. Tan, X. X. Yang, and B. Han, "A dual-polarized Fabry-Perot cavity antenna at millimeter wave band with high gain," in *2015 IEEE 4th Asia-Pacific Conference on Antennas and Propagation (APCAP)*, pp. 621-622, 2015.
[12] H. Aliakbari, A. Abdipour, R. Mirzavand, A. Costanzo, and P. Mousavi, "A single feed dual-band circularly polarized millimeter-wave antenna for 5G communication," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*, pp. 1-5, 2016.
[13] S. Hur, S. Baek, B. Kim, Y. Chang, A. F. Molisch, T. S. Rappaport, K. Haneda, and J. Park, "Proposal on Millimeter-Wave Channel Modeling for 5G Cellular System," *IEEE Journal of Selected Topics in Signal Processing,* vol. 10, pp. 454-469, 2016.
[14] C. J. Hansen, "WiGiG: Multi-gigabit wireless communications in the 60 GHz band," *IEEE Wireless Communications,* vol. 18, pp. 6-7, 2011.
[15] W. H. Yang J., Lv Z., Wang H., "Design of miniaturized dual-band microstrip antenna for WLAN application," *Sensors,* vol. 16, pp. 1-15, 2016.
[16] C. A. Balanis, *Advanced Engineering Electromagnetics*, Second Edition ed.: JohnWiley & Sons, New York, 2012.
[17] (2015) CST Microwave Studio. Available: https://www.cst.com/products/cstmws
[18] High Frequency Surface Structure (HFSS) (15 ed.). Available: http://www.ansys.com

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 501-514 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Building an Online Interactive 3D Virtual World for AquaFlux and Epsilon

Omar Al Hashimi*, Perry Xiao

London South Bank University, School of Engineering, SE1 0AE, UK

ARTICLE INFO

Article history:
Received: 31 October, 2018
Accepted: 16 December, 2018
Online: 23 December, 2018

Keywords:
3D modelling
Virtual Reality(VR)
AquaFlux and Epsilon
3ds Max
Web 3D applications
Virtual User Manual (VUM)

ABSTRACT

In today's technology, 3D presentation is vital in conveying a realist and comprehensive understanding of a specific notion or demonstrating certain functionality for a specific device or tool, especially on the World Wide Web. Therefore, the importance of this field and how its continuous enhancement has become one of the dominant topics in web development research. Virtual Reality (VR) combined with the use of a 3D scene and 3D content is one of the best delivering mechanisms of this realist ambience to users. AquaFlux and Epsilon are clinical instruments that were built, designed, and developed at London South Bank University as research projects for medical and cosmetic purposes. Currently, They have been marketed and used in almost 200 institutions internationally. Nevertheless, considering the type of these tools, they often involve on-site thorough training, which is costly and time-consuming. There is a real necessity for a system or an application where the features and functionalities of these two instruments can be illustrated and comprehensively explained to clients or users. Virtual User Manual (VUM) environment would serve this purpose efficiently, especially if it is introduced in 3D content. The newly created system consists of a detailed virtual guide that will assist users and direct them on how to use these two devices step-by-step. Presenting this work in a VR immersed environment will benefit clients, user and trainees to fully understand all the features and characteristics of AquaFlux and Epsilon and to master all their functionalities.

1. Introduction

This paper is an extension of work originally presented in *Advances in Science and Engineering Technology International Conferences (ASET) 2018* [1]. The current research paper presents the development process of a web-based interactive 3D virtual world that illustrates and covers all steps of how AquaFlux and Epsilon operate. VR collectively used with 3D objects presentation will efficiently serve the purpose of illustrating all features and functions of any newly purchased device in an immersed world that gives viewers a real feel of the experience. The popular change to VR in online training or education is going to promote a new learning concept, where clients, trainees or even students not only gain knowledge but also communicate with each other by changing content in a variety of ways. The main feature of VR is the prospect of social interaction, providing the ability for immediate actions and reactions in real time. VR environment has constantly been associated with 3D modelling; it is by far one of the best ways to illustrate and show any 3D content, object, scene, model, etc. The word virtualisation generally depicts the separation of a resource or request for a service from the underlying physical delivery of that service. An additional factor that makes virtualisation very practical and useful is the interactivity that 3D multimedia applications can provide it, particularly when the whole project displayed on-line using the WWW [2]. Moreover, users of the current Internet age are ready to shift from 2D online presentations to 3D. The arrival and the broad use of 3D web materials have amplified the necessity to acquire improved 3D technologies and generate an extremely advanced and practical product. Therefore, it is suitable to upgrade and improve the usage of those 3D technologies and interactive environments to new fields, like e-learning, Virtual Learning Environment (VLE), museums, e-commerce, online training, Learning Management System (LMS), tourism, health and the government part. Adapting these enhanced 3D virtual interactive apparatus into those latest fields has immensely improved and upgraded the user understanding and became stimulating. Online 3D VR worlds and contents have amplified users' perceptions as they deliver and comprise physical world appearances and attributes to users and permit them to connect to it resulting in immersing users into their environments. 3D VR world, used through the WWW, would be as viable as using an application locally. Therefore, the concept of VR can be incorporated with social media to additionally increase its attraction. Nonetheless, a new VMware report explains that the promising future of 3D/VR technology in our daily applications

*Omar Al Hashimi, London South Bank University, alhashio@lsbu.ac.uk

can be entirely used solely if joined with the improvement of a competent and simple to use ways of building, controlling, search and demonstration of interactive 3D multimedia content, that can be practised by skilled and novice users [3]. In this context, London South Bank University's engineering lab has created and designed AquaFlux and Epsilon, which are clinical instruments. Displaying these medical devices in the VR approach would demonstrate their capabilities as being practised in the real world. Xu (2015) states AquaFlux as:

> *a new condenser based, closed chamber technology for measuring water vapour flux density from arbitrary surfaces, including in-vivo measurements of transepidermal water loss (TEWL), skin surface water loss (SSWL) and perspiration. It uses a cylindrical measurement chamber [4].*

Compared to other technologies, AquaFlux has greater sensitivity, greater repeatability, and most importantly, the measurement results are independent of the external environment. Biox (2014) describes Epsilon as:

> *a new instrument for imaging dielectric permittivity (Ɛ) of a wide variety of soft materials, including animal and plant tissues, waxes, fats, gels, liquids, and powders. Its proprietary electronics and signal processing transform the sensor's native non-linear signals into a calibrated permittivity scale for imaging properties like hydration. [5].*

The Epsilon user manual describes that "The system consists of a hand-held probe, a parking base, and an in-vitro stand, securely stored in a purpose-designed case" [6]. Both devices are shown in Figure 1.

Figure. 1. Medical instruments AquaFlux (L), Epsilon (R)

2. The utilisation of 3D contents and related work

After the advent of Web 2.0, there has been a major improvement in web applications. Web 2.0 supports and encompasses various functions, like team working, communicating, and the connection between computer and Internet users. 3D immersive virtual worlds (3DVW) are one of the important applications of Web 2.0, which are computer-generated, virtual, online, graphics, multimedia and 3D worlds [7]. The major notion of building such an object in 3D is to illustrate to users the reality of the content. 3D modelling is an essential element in the area of VR technology [8]. The development of a 3D content can be divided into three main steps:

1- 3D modelling
2- Layout and animation
3- Rendering

Normally, 3D models are created using 3D modelling software, like 3ds Max, or Maya, or their open source equivalents – although open source tends to be less complex with fewer advanced features. A 3D scene model is created from geometrical shape objects, for example, rectangles, triangles, circles, cones, etc. e-learning contents are placed into hypermedia documents, causing difficulty in 3D integration into HTML files. Web browsers are not yet designed to deal with the (normally large) 3D data files which require a large number of computational resources. It takes time and effort to digest such proposals. Therefore, readily available Mozilla and Google proposals can be adopted as standards. Mozilla and Google use the same technologies: OpenGL interfaced with JavaScript [9].

3D objects are widely used in science, technology, engineering, health, education, cosmetics, simulation, e-learning etc. It has been used in the areas named above joined with VR to stimulate the displayed content as it is applied and practised in the physical world. Another aspect that 3D contents and VR play a very active role in, is the Vocational Education System. It is a system that is specifically designed to support the industry and manufactures by offering vocational and technical courses delivered in VR ambience and PLE (Personal Learning Environment) which allows users, learners to control their learning and manage their own learning experience (distance learning). Kotsilieris, Dimopoulou (2013) point out:

> *The development of 3D Virtual Worlds plays an important role in e-learning and distance learning. Through three main features: I) creating the illusion of a 3D environment, II) support the application of avatars as virtual representations of human users, III) offer communication and interaction tools to their users. The evolution of virtual worlds is a result of a rapidly evolving field of electronic games. In brief, Virtual Worlds are designed to offer real-time communication tools, interaction capabilities and collaboration [10].*

Some up-and-coming technologies will overpower some difficulties that are faced currently in areas like education, technology, health, engineering and others. These include computer graphics, augmented reality, computational dynamics and virtual worlds. Lately, we have witnessed a number of novel thoughts emerging in the literature related to the future of education. As has been pointed out by Potkonjak, Gardner, Callaghan, Mattila, Guetl, Petrović, Jovanović (2016):

> *Technological examples mostly related are e-learning, virtual laboratories, VR and virtual worlds [11].*

Accordingly, the benefit of adopting VR in the medical and healthcare sectors is to teach and train medical students, trainees and clients on how to use medical devices and instruments and how to conduct some medical tasks. Web-based and online 3D objects used in medical training tools and environments showed to develop the educational process. This web-based virtual medical system of devices for cardiac diagnostic and monitoring functionalities has been created and built to assist in the process of training medical students, qualified health personnel and non-medical staff to carry out an Electrocardiogram (ECG), an Automatic External Defibrillator (AED) and a blood pressure device [12].

Those applications guarantee an interactive e-learning experience in the medical field. Also, considering the main objective to emulate real patients, anatomic regions, and clinical

tasks and to represent real-life conditions in which this medical tool is built for. These virtual environments allow interaction between users and the system as well as manipulation with very sensitive reactions similar to that of real-life objects. This type of system will promote learning by practising, which makes the whole experience straightforward and enjoyable [12]. Using VR in such projects will help to achieve an extremely immersive experience [13].

Another related work in this regards is the ARCO project (Augmented Representation of Cultural Objects). The ARCO project is specifically built for the tourism and heritage industry. Its main objective is to build up an entire virtual museum that contains a collection of technologies for producing, changing, controlling and demonstrating cultural objects in VR environment that are accessible globally via the Internet [14].

3. The implementation of the interactive Virtual User Manual (VUM) of the medical instruments

The design process of AquaFlux and Epsilon medical devices was slightly difficult and experienced the array of choices for selecting the suitable 3D modelling software; in this project, the software modelling tool used is 3D Studio Max (3ds Max). In addition, adding interactivity between users (clients) and the devices, the software Adobe Flash CS6 is used. Other modelling software products were used at the early stages of this project, like Google Sketchup, Blender and Unity were used as a final result of the research was to create, build and develop the objects and all 3D scenes in 3ds Max for its professionalism and the complete collection of functions and options that it had to offer.

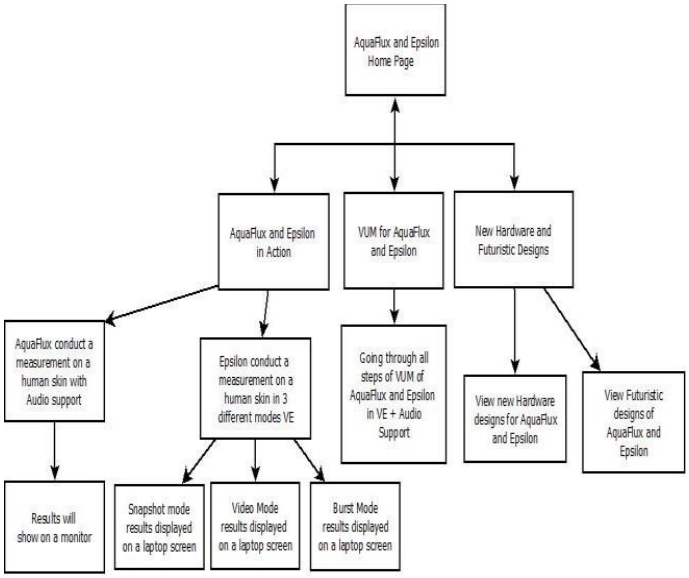


Figure. 2. Flowchart of AquaFlux and Epsilon online 3D Environment

AquaFlux and Epsilon are clinical devices used for skin treatment, belonging to the health and medical sector. Figure 2 shows the workflow diagram that illustrates all the steps of the AquaFlux and Epsilon 3D virtual environment system and describes the development processes of the Virtual User Manual (VUM). At the home page, users have three options: AquaFlux and Epsilon in Action, users can perform virtual skin measurements using the instruments. For VUM of AquaFlux and Epsilon, users can go through all the training steps, with audio and illustrative text support. For New Hardware and Futuristics Designs, users can view the new hardware designs and futuristic concepts for both medical instruments.

The health sector is largely connected with the utilisation of leading technologies like 3-dimensional objects and virtual presentation for demonstration, education, testing and performing a number of medical operations and routines. 3D Virtual Worlds (3DVWs) have been used in a variety of applications in the medical sector and applied in health-related activities [7].

3.1. The development and design of AquaFlux and Epsilon using 3ds Max modelling software

3ds Max modelling tool can swiftly and professionally produce 3D scenes and objects [15]. Like in various 3D development tools, initially, we have to create and design the medical tools. AquaFlux and Epsilon contain a base part and a probe; on the base part, there are little design obstacles, e.g. cabling input port and buttons that have to be switched left or right in AquaFlux's case. The probe and the base have been calculated with an actual ruler to let us build the models in an extremely realistic form. Additionally, considering some photos to get an accurate proportion and size for every model, for instance, how large the probe would be in comparison to a human hand.

Figure. 3. AquaFlux probe measured by a ruler

Once the developer assesses the accurate size of every object, we use 3ds Max to design the medical devices. Modelling software is shown in the figures below for AquaFlux and Epsilon:

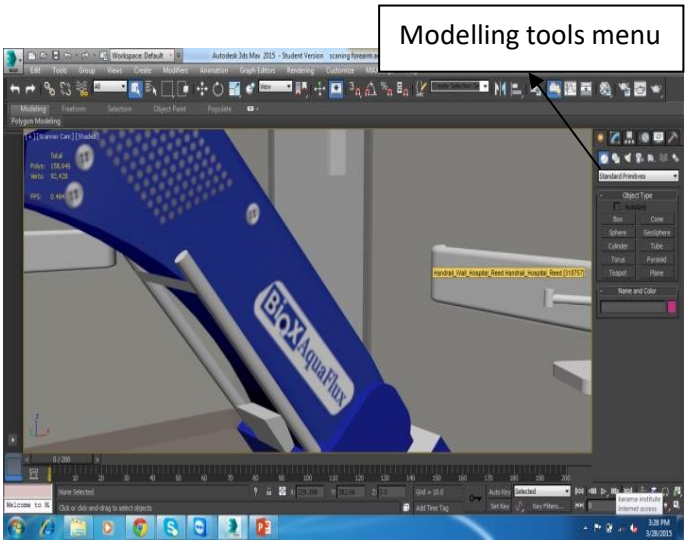


Figure. 4. Tools box menu 3ds Max software used to model objects

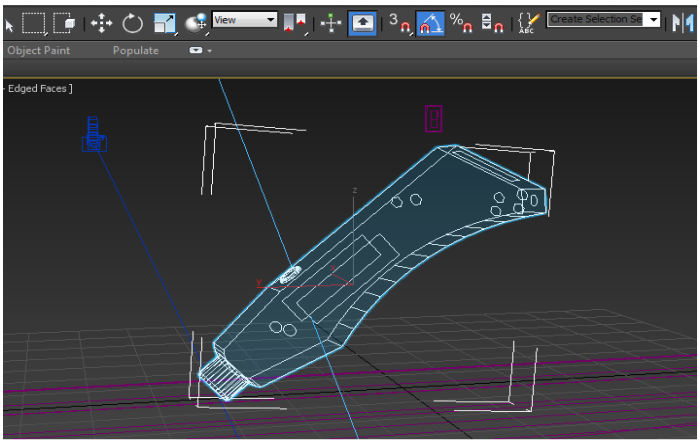


Figure. 5. AquaFlux probe modelling in 3ds Max

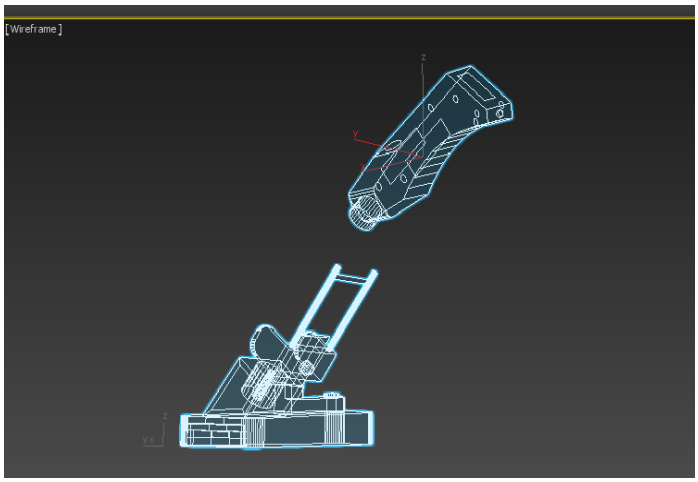


Figure. 6. Snapshot of AquaFlux probe and base design in 3ds Max

The designer is able to choose a diverse object's shapes from the modelling menu e.g. a box, cone, and sphere, tube and cylinder and gives names for all parts. As the process of modelling is progressing, the developer has to construct a real reference to check how the recently created model is comparable to the real object via situating the real model's image close to the newly designed one as shown in Fig 8:

3.2. Adding materials to AquaFlux and Epsilon

Materials and colours are considered early to befit the model reference if it is really sufficient to satisfy the client's demands. The selection process of the material is closely examined to achieve a minimum rendering time and to present a very real product. Materials that are rich in glossiness and reflection will consume most of the processing (rendering) time, however, without materials, the model would be very unreal to viewers. Vray materials are used in this project, and the third-party renderer is Vray for the whole work. Vray is one of the industry standards for generating specialised pictures. For the Epsilon Probe and other objects, a semi-gloss with a whitish material is used with 0.6 glossiness and a subdivision of 20. Those settings were chosen to reduce the time it usually takes during the rendering process, as rendering is a crucial part of any 3D creation projects.

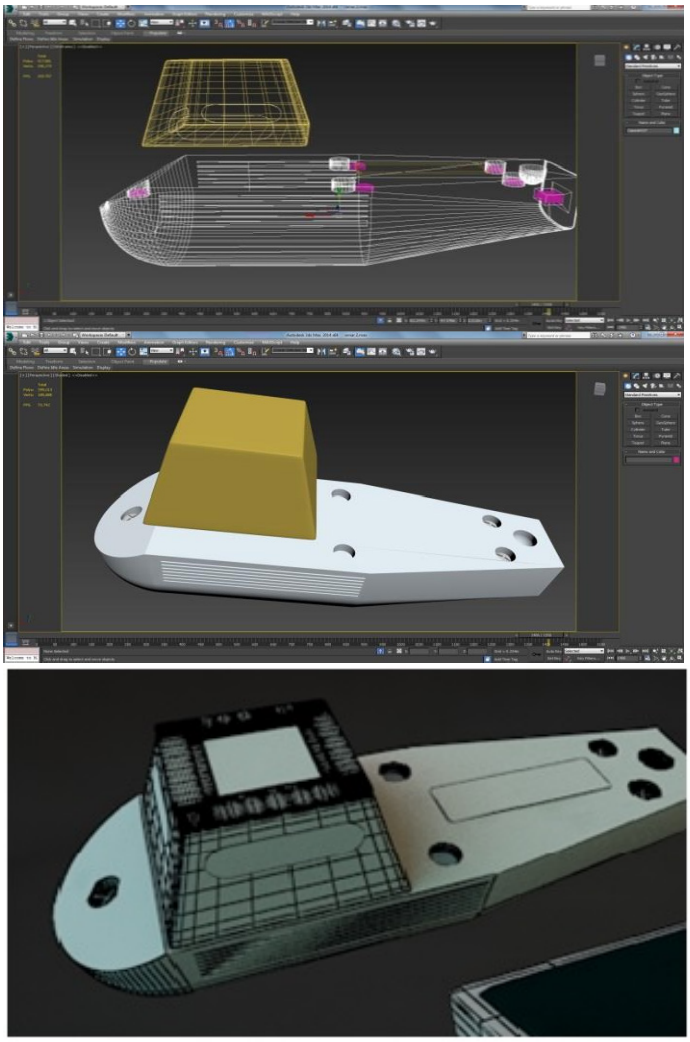


Figure. 7. Different stages of modelling Epsilon's probe in 3ds Max

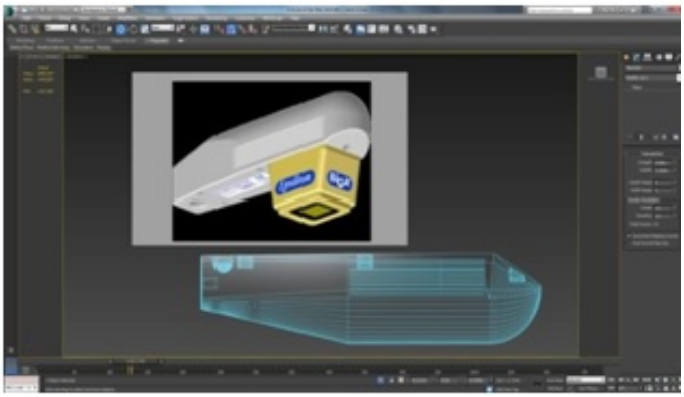


Figure. 8. The original picture of Epsilon compared to an Epsilon object.

In this project, High Definition Range Imaging (HDRI) is used for the background to make it slightly more practical for the lights and shadows. HDRI is absolutely valuable to emulate shadows and lightings of a real environment that will be projected on the objects in the 3D scene. Although using HDRI will add some extra time during rendering each scene, the glossiness and reflectivity of the model are changed slightly low, to improve the overall execution-time process.

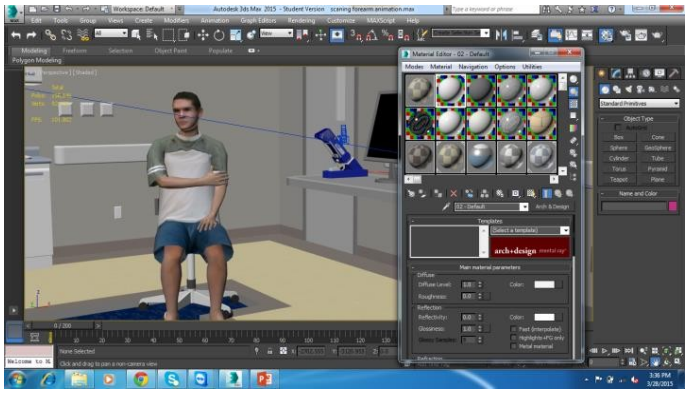


Figure. 9. Colours added to AquaFlux using the material box menu

Figure. 10. Materials menu in 3sd Max (L), Materials added to Epsilon probe (R)

Once the materials eventually determined and added taking into account the time and quality of the work, we arrange all 3D contents for the next process: Animation. In order for all the objects in our 3D scene to be animated correctly, all objects in the selected scene must be grouped together via animation attach buttons in the command panel under edit geometry menu. This means all objects should be grouped, attached or sometimes called fused together. If any object, for any reason, has broken up and not been linked and left out of the attached objects group, the keyframes associated with that model will disappear after this stage and the model will no more be animated.

3.3. The process of animating AquaFlux and Epsilon

Animating 3D objects is an essential phase of our project's design. Almost a third of the time of this project was used on the animation process of our objects that were modelled and built earlier. Planning is critical at this stage for each planned scenario. Planning should be prepared and anticipated for every object in our scene: how the object will interact with other objects, the proposed movement and the path of that motion that the individual object will experience in the virtual world. It is a somewhat prolonged process. Contrary to character's animation, this type of animation is slightly easier as it does not require bones and additional complicated rigging tools. Animating objects in this work has been accomplished in two ways: attaching an object into a group of objects and using the collapse utility to cave in multiple selected objects into a single object. The necessary points for all movements must be thoroughly calculated prior to the animation phase. In the animation stage, the Timeline is set to 1000 frames and 15 frames per second because this work is specifically designed for online environment only. The number of frames per second is decreased to suit the file size of the images that will be rendered to be used on the WWW. Applying curve editor menu options for animating all objects and scenes. When modifying curve points, all animated models can be managed via (speed, timing, and movements options).

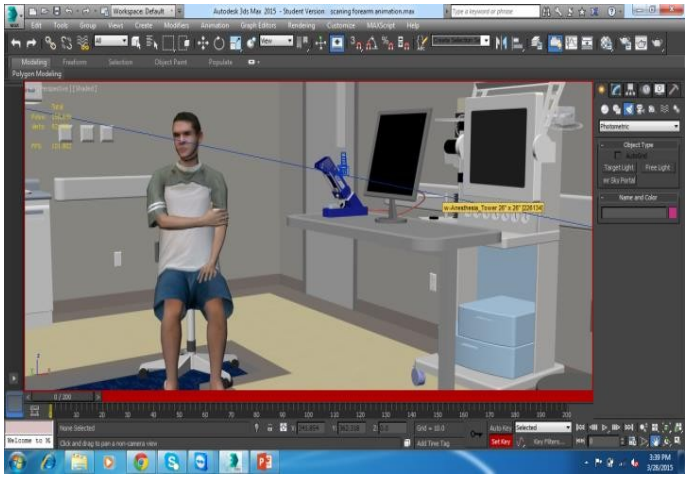


Figure. 11. Adding keyframes (redline) in 3ds Max

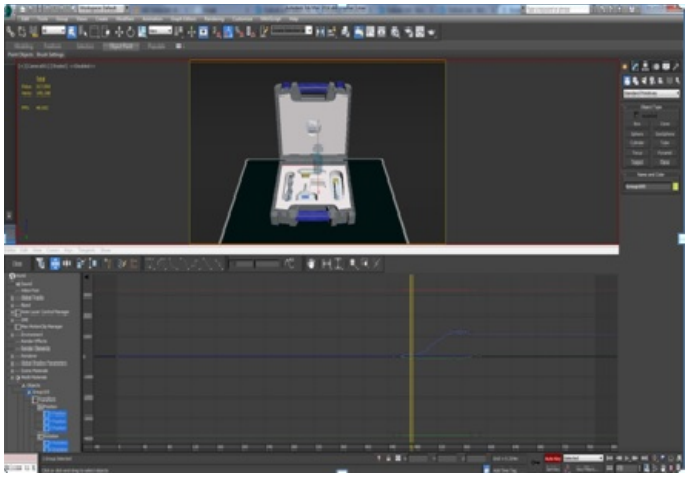


Figure. 12. Using the Curve editor in 3ds Max

3.4. The process of rendering AquaFlux and Epsilon

After the animation is completed and setting keyframes from a specific position into where we want our object to stop at a certain point. Currently, all objects are now prepared for the last phase of the modelling process, which is rendering. In rendering the developer has to monitor all frames being rendered. Nevertheless, at this point in our work, it is extremely vital to monitor the speed and time that each object takes to complete rendering. Every frame is closely checked if it is in the correct

position (number) and whether all models have the exact shadows and lighting. Occasionally models are spotted floating that are unnoticed in the animation process. In rendering, a small window will pop up showing how rendering each square pixel of the object, calculating the objects materials, colours, shadows, light reflections, etc. In the rendering options box, it is possible to select the format of the rendered images and their file locations that will be exported into Adobe Flash CS6 later on for adding interactivity via linking all rendered images, scenes and objects.

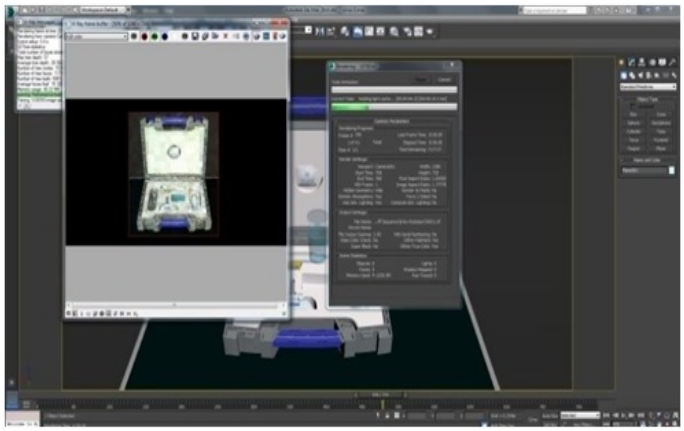


Figure. 13. Rendering window after all keyframes and files setup completed

3.5. Creating links and adding interactivity to AquaFlux and Epsilon

All rendered image sequences of the medical devices AquaFlux and Epsilon have to be exported from 3ds Max software into Adobe Flash CS6 in order to add interactivity to the project via creating interactive buttons.

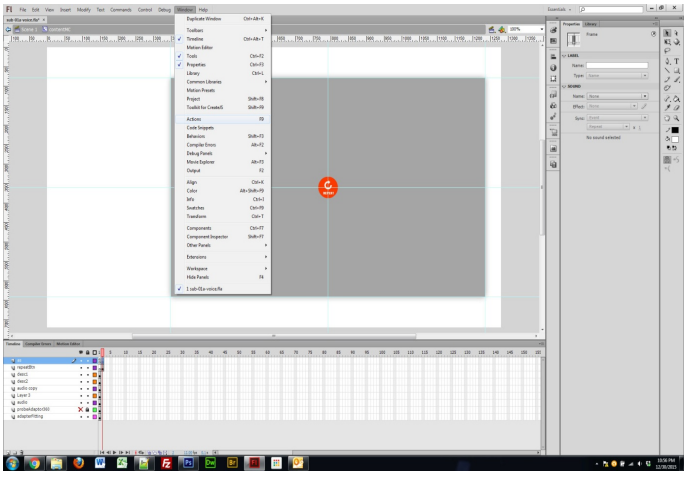


Figure. 14. Adding ActionScript into a scene in Flash CS6

AquaFlux VUM is a completely interactive website that demonstrates all functions, features and a sample skin measurement process conducted on a human.

The AquaFlux VUM is an interactive system. Users can click on the device's probe to command it to go towards a human's forearm to perform a sample skin measurement. Each step of the VUM is assisted with audio instructions recorded earlier to assist in exemplifying the operational process for this medical device.

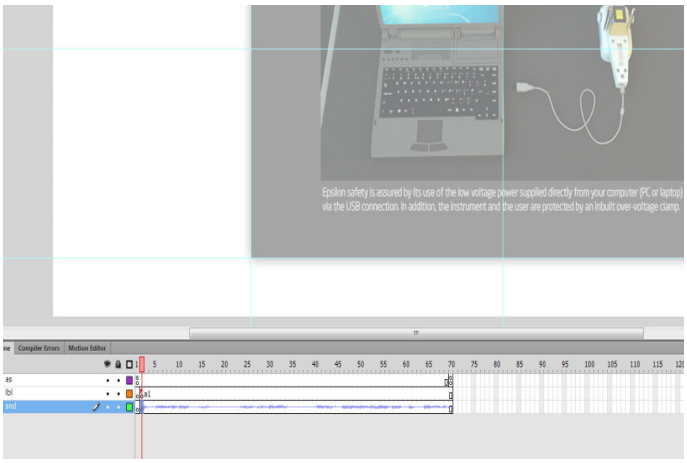
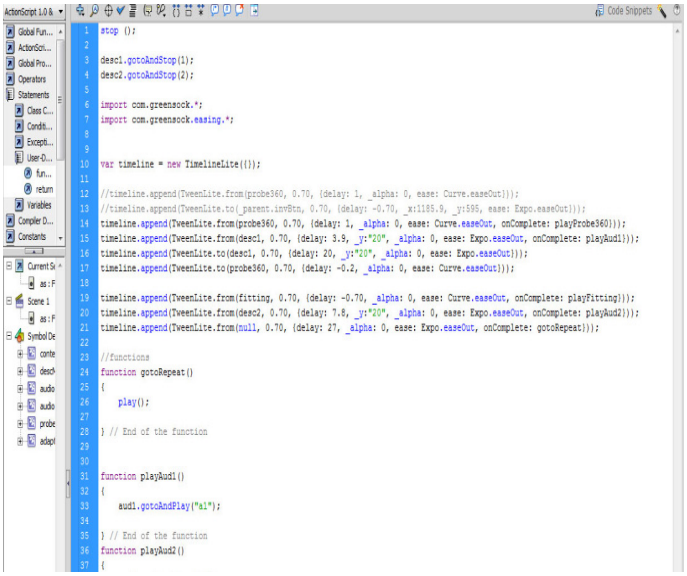


Figure. 15. Syncing a voice clip with the text frame in Epsilon scene



```
stop ();

desc1.gotoAndStop(1);
desc2.gotoAndStop(2);

import com.greensock.*;
import com.greensock.easing.*;


var timeline = new TimelineLite({});

//timeline.append(TweenLite.from(probe360, 0.70, {delay: 1, _alpha: 0, ease: Curve.easeOut}));
//timeline.append(TweenLite.to(_parent.invBtn, 0.70, {delay: -0.70, _x:1185.9, _y:595, ease: Expo.easeOut}));
timeline.append(TweenLite.from(probe360, 0.70, {delay: 1, _alpha: 0, ease: Curve.easeOut, onComplete: playProbe360}));
timeline.append(TweenLite.from(desc1, 0.70, {delay: 3.9, _y:"20", _alpha: 0, ease: Expo.easeOut, onComplete: playAud1}));
timeline.append(TweenLite.to(desc1, 0.70, {delay: 20, _y:"20", _alpha: 0, ease: Expo.easeOut}));
timeline.append(TweenLite.to(probe360, 0.70, {delay: -0.2, _alpha: 0, ease: Curve.easeOut}));

timeline.append(TweenLite.from(fitting, 0.70, {delay: -0.70, _alpha: 0, ease: Curve.easeOut, onComplete: playFitting}));
timeline.append(TweenLite.from(desc2, 0.70, {delay: 7.8, _y:"20", _alpha: 0, ease: Expo.easeOut, onComplete: playAud2}));
timeline.append(TweenLite.from(null, 0.70, {delay: 27, _alpha: 0, ease: Expo.easeOut, onComplete: gotoRepeat}));

//functions
function gotoRepeat()
{
    play();

} // End of the function


function playAud1()
{
    aud1.gotoAndPlay("a1");

} // End of the function
function playAud2()
{
```

Figure. 16. ActionScript code for linking the button to play, replay and stop the audio clip

The above steps (Figures 14, 15 and 16) can be repeated to all the 3D scenes and objects that required some form of connectivity and interactivity with the user in a VR atmosphere by using buttons in Flash CS6.

Figure. 17: The result of the process of skin measuring will be shown on a screen

Epsilon VUM follows a similar approach to conducting a human skin measurement procedure for medical purposes. Figure 18 shows the Epsilon instrument in action.

Figure. 18. Skin measurement process taking place on a patient using Epsilon

Once carrying out the measuring process, the scene's camera will move towards the laptop's screen to display the scan results as shown in Figure 19.

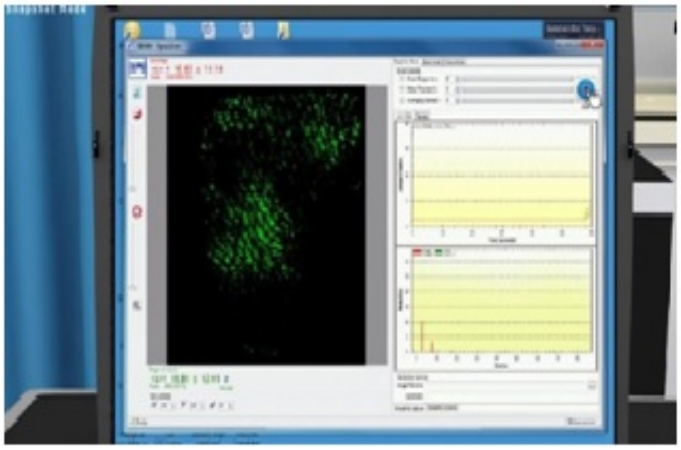


Figure. 19. The result of the Epsilon scanning process is displayed once clicked on the blue button, top right corner.

Figure. 20. Epsilon protective case and all its components are shown in a 3-dimensional interactive atmosphere

In Epsilon VUM, users are able to perform human skin scanning in three various approaches, snapshot, burst and video mode. All results of scans (images and videos) are stored in a folder on the laptop's hard disc. Scan buttons and tabs are all interactive.

4. Designing and introducing new accessory, hardware and holders of AquaFlux and Epsilon

During the final stages of this project, a fresh design plan has evolved, which will contribute positively to the entire skin measurement process. Moreover, it will make a great addition to the practicality of the two devices. The notion was to develop and build a novel part (holder) for both tools. It will contribute massively to the benefit of the entire process making it more effective and competent by allowing the simultaneous measurement of several patients. The newly designed holder for Epsilon was thoroughly considered and particularly built to suit the Epsilon structure and patient's easement. The material added to the newly developed holder was a cloth semblance material to emulate the real world object.

Figure. 21. Epsilon newly designed holder strap in action

In relation to Aquaflux, it was rather sensitive to create the novel holder due to the shaped probe. After a thorough analysis of the structure of the device, AquaFlux's new accessory was completed. The new holder already considered both patients relive and client's ease. AquaFlux probe is made of metal and fairly heavier in comparison to Epsilon. Thus, the black frame around the border of the holder is recommended to be manufactured of the magnet material providing AquaFlux with further support, firm placing, and accurate reading.

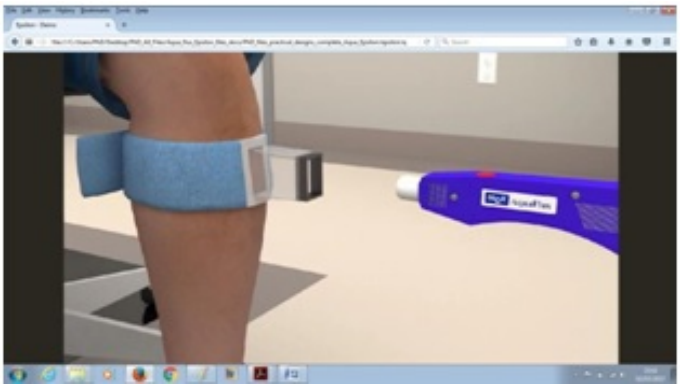


Figure. 22. AquaFlux newly designed holder strap with magnet head

Conclusively, the novel ideas of designing and building new hardware for these medical tools has optimised the functionality of both devices. Furthermore, to provide a broad number of people concurrently with the service, futuristic concepts and designs have evolved. The innovative designs of AquaFlux allow it to rotate in all directions giving it extra flexibility and ease of use for all parties involved in the process as shown in Figure 23 below:

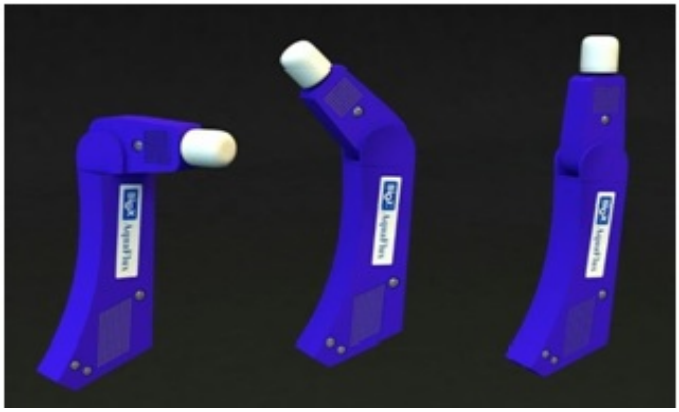


Figure. 23. Rotatable AquaFlux for further flexible positions

The AquaFlux arm-band is an entirely new idea, extremely lightweight and it allows further ease of use. It is padded and soft which can be used almost everywhere on the human's body. It has a Bluetooth capability to connect to a PC or laptop wirelessly and spare the trouble of cabling that can be distracting.

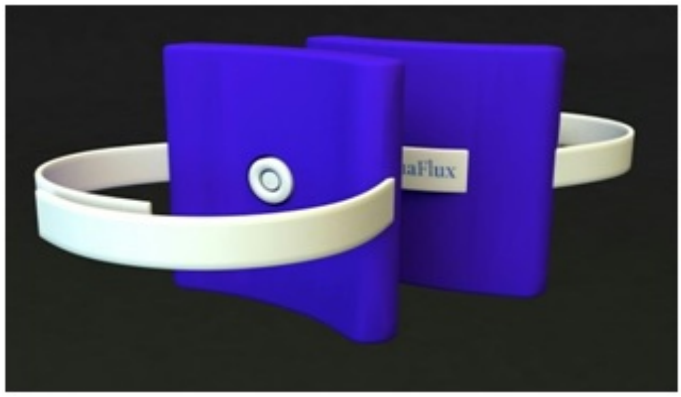


Figure. 24. AquaFlux wireless stripe, newly designed hardware

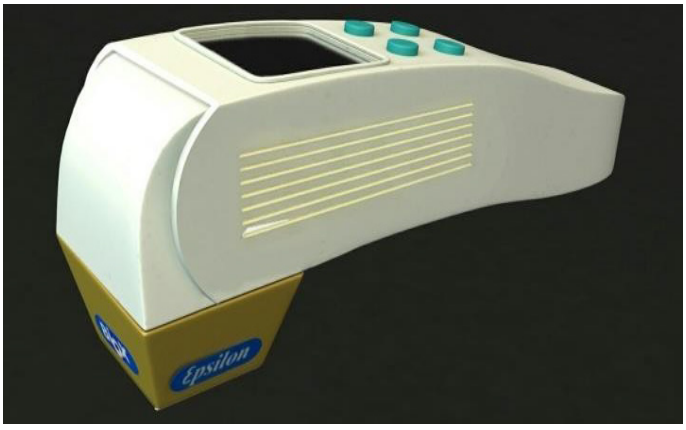


Figure. 25. Epsilon turning head, on probe monitor and operational buttons

Epsilon's turning head assists to carry out skin measurements in vertical/horizontal positions. A fresh screen is supplemented to the initial probe for the purposes of accessibility, simplicity and comfort. Additionally, Functional buttons have been supplemented to the top surface of the handle to offer on probe actions, e.g. on/off, scan, reset. Epsilon device can be wholly moveable.

5. Virtual Reality creation

Virtual Reality is a second world (simulated environment) accessed by various users, and it resembles our world in most of its characteristics depending on what is required from that virtual world to present. Users in virtual worlds can be presented by avatars or walkthrough scenes to conduct a variety of activities.

There are currently two methods to create a VR environment:

1- Using a 360-degree video: this method is easier than the second one, as you can record real-life video footages and use those footages in building the content. This method needs another tool at the later stages to view the 3D content, HMD (Head Mounted Display) unit. With this HMD unit, users can experience the new world of computer simulated reality that imitates physical existence in the actual world.

2- Using 3D animation: this method is when all selected objects will be created and developed in 3D modelling tool e.g. 3ds Max, Blender, Maya and many more. It is more expensive as it requires more tools and software to be involved in the construction process as well as requiring the expertise to develop, create, animate, render and subsequently publish the product online. Another way of creating VR is by using game engine software such as Unity. This method is mainly used for creating and developing games.

From the two methods mentioned above, it is clear that the second method does require a huge amount of effort and experience. All aspects of the second method need to be carefully and professionally designed and planned. No real images or footages that can help the creator to replace any physical existence in the virtual world; all have to be created, built and designed from scratch. In addition, it is the method that has been adopted in this work.

Figure. 26. AquaFlux interactive user manual website contains three links

6. Results

6.1. AquaFlux Interactive User Manual

AquaFlux VUM online system is a completely interactive website that displays a sample skin measurements conducted on a human and illustrates all functions and features of the medical tool. The following Figures 26, 27, 28, 29, 30, 31, 32 and 33 present few steps on how to use the VUM.

By clicking on the button labelled Aquaflux in Action, will present a sample of skin measuring process conducted on a client at a clinic.

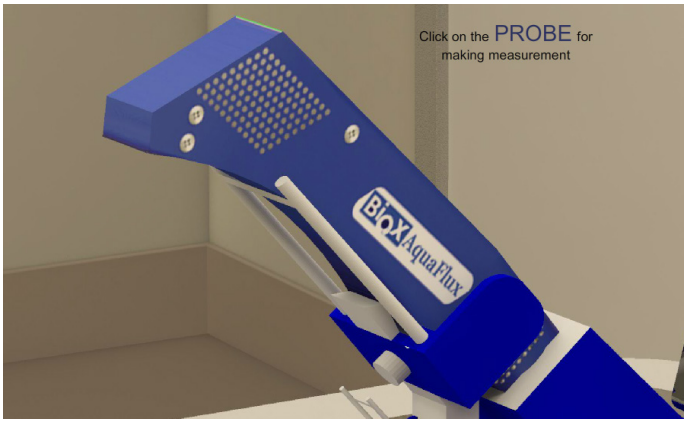


Figure. 27. AquaFlux in its base, waiting for a user to click the link above it

Figure. 28. AquaFlux in action, moving to a patient's forearm for a measurement

After clicking on the link in Figure 28, the AquaFlux probe will move towards a patient's forearm to conduct a sample skin measurement, by placing its head that has a cap into the patient's skin.

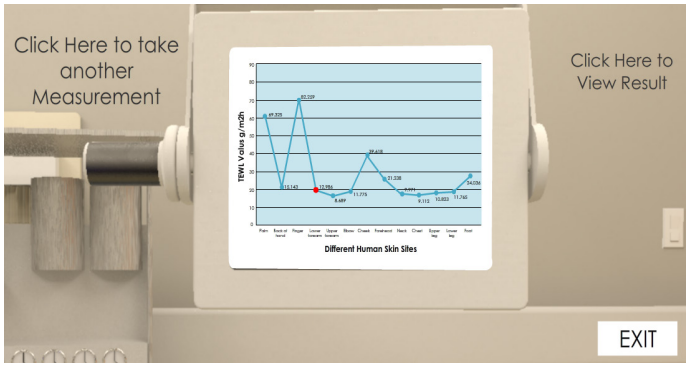


Figure. 29. AquaFlux monitor shows measurement results and other user's options

Once clicked on view results link (top right corner), the reading will show on the screen with an accurate reading of the forearm. A user, a trainee or a client also has another option to take another measurement (top left) or to exit to the main page via the exit button (bottom right).

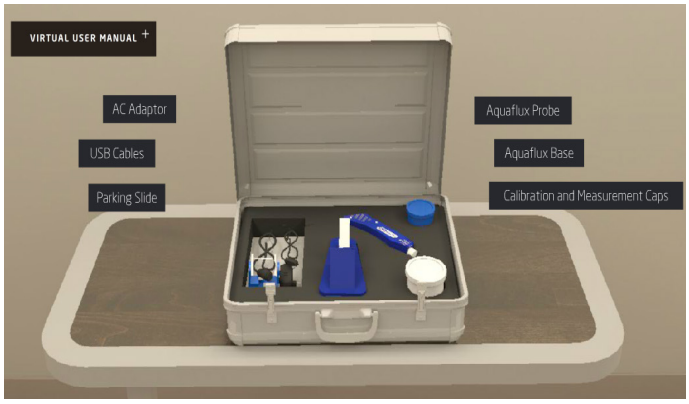


Figure. 30. The 2nd option link to view the VUM

From the main page, clicking the second option takes the user into the AquaFlux VUM page. Here, users can click on the VUM (top right) link to view AquaFlux setting up the process, calibration, caps, AC adaptors, connectors, probe parking, holding the probe, software familiarisation. Also, the process of how a user can fit and slide the probe into the parking base and an explanation of the AquaFlux rear panel unit.

Figure. 31. AquaFlux 3D components illustrated in the protective case

As shown previously in Figure 31, once the mouse is placed over one of the other six links, it will popup the part and rotate it in 3D mode; this gives a chance to the user to know all the right names for the AquaFlux device components.

Figure. 32. Clicking on the VUM menu, the background becomes blurry

In Figure 38, users can click on VUM's options' menu to display all list items of AquaFlux components. The menu is designed to illustrate all parts of the medical device as well as demonstrating the role of each part in the measuring process displayed in a 3D environment.

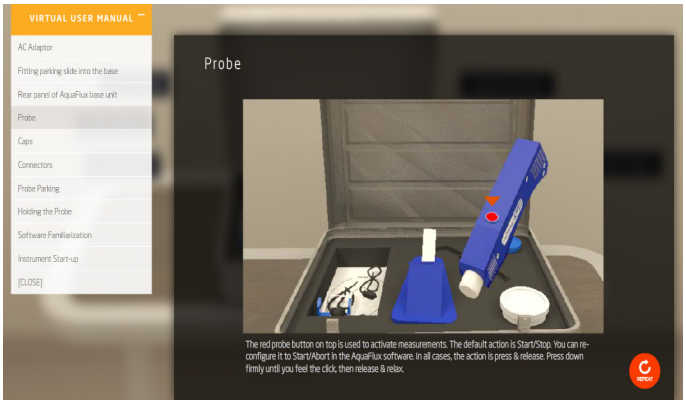


Figure. 33. Clicking on one of the options in the VUM menu (probe).

In Figure 33, the user clicked on probe item in the VUM's drop options box, a window appeared to the right of the screen containing instructions and info regarding the AquaFlux's probe with brief descriptions at the bottom of the page. Users can start again the step by clicking on the repeat button (bottom right corner). To go back, users need to click on the [CLOSE] link first then clicking on the link back to the home page.

Figure. 34. The main page of Epsilon VUM contains three main links

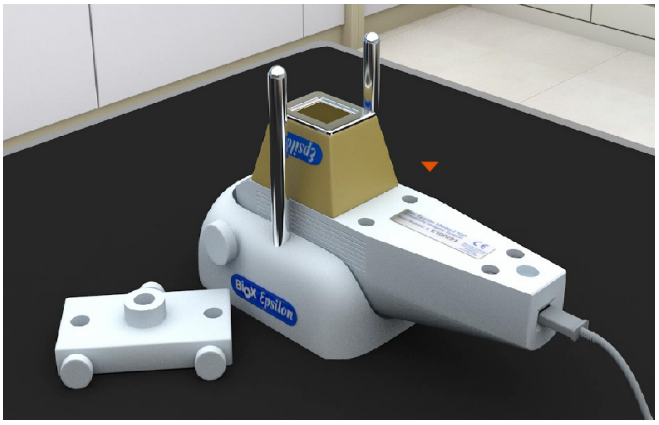


Figure. 35. Clicking on the 1st link to see Epsilon in action

6.2. Epsilon Interactive User Manual

Epsilon VUM online system is a completely interactive website that displays a sample skin measurement conducted on a human and shows all functions and features of the device.

After clicking on the Epsilon probe (see Figure 35) the probe will move towards the patient's hand and placed there. Then the scene will switch into the laptop. After that, the camera moves towards the screen to see the result of skin measurement taken. All the above steps are conducted with the assistance of audio feature.

Figure. 36. Epsilon probe moving towards the patient's hand for a measurement

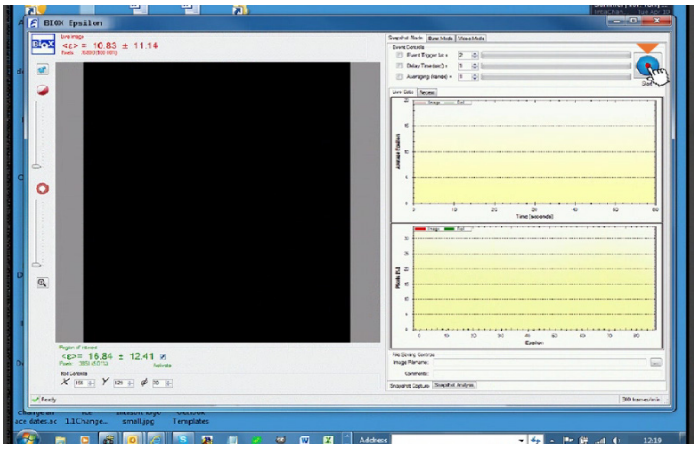


Figure. 37. Clicking on the button (top right) to start scanning human skin

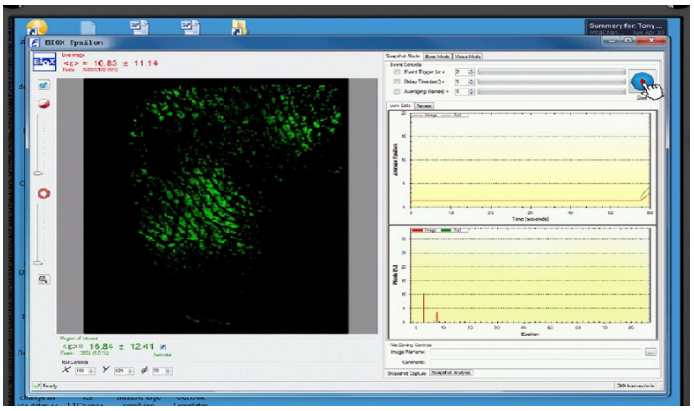


Figure. 38. Scanning in a snapshot mode

After clicking on the scan button, scanning will begin in a snapshot mode, in Epsilon users can scan in three various approaches (snapshot mode, burst mode and video mode).

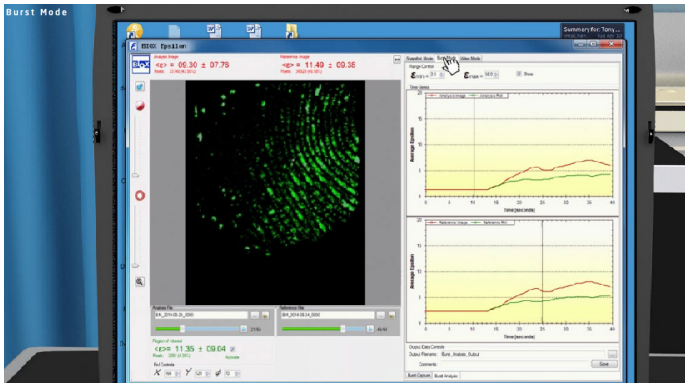


Figure. 39. Scanning in a burst mode (clicking the tab at the top to switch mode)

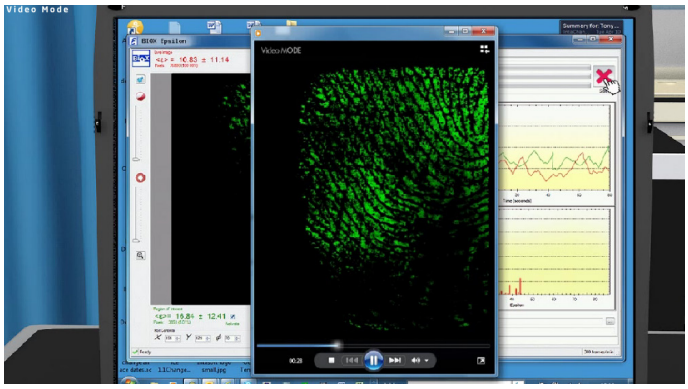


Figure. 40. Scanning in a video mode

All files, images and videos of the three different scanning modes saved in their respective folders, snapshot, burst and video folders, users can go to their folders and view, replay or send the captured files. This VUM is assisted with an audio feature to clarify any unclear steps throughout the demonstration process.

Figure. 41. Epsilon 3D components in a sturdy case

As previously explained in the AquaFlux VUM (Section 7.1), the second link in Epsilon's main page as shown in Figure 34, will take the user to the Epsilon's VUM menu. Every part or component located in the protective case will pop up once the user places the mouse on it and start rotating in 3D 360 degree with a brief thumbnail defining that part. The use of mixed technologies of 3D designs and VR results in achieving an incredible engulfing feeling as if being at the real location of the demonstration process [13].

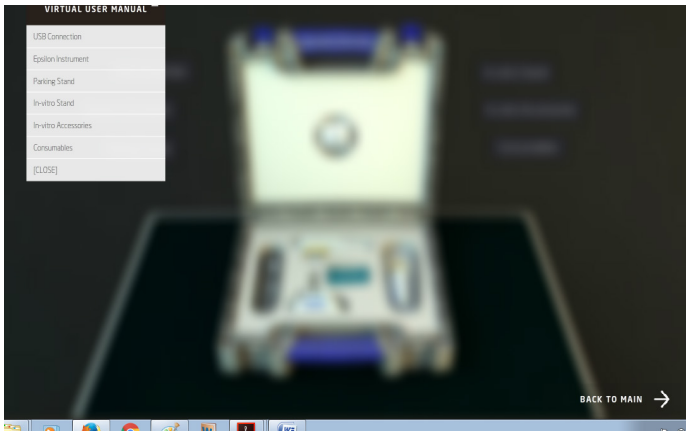


Figure. 42. Epsilon VUM drop down menu

In a similar manner to the AquaFlux pervious VUM, Epsilon's VUM contains interactive links to demonstrate USB connection, Epsilon instrument, parking stand, in-vitro stand, in-vitro accessories and consumables.

Figure. 43. Epsilon probe sliding into the stand in the VUM menu options

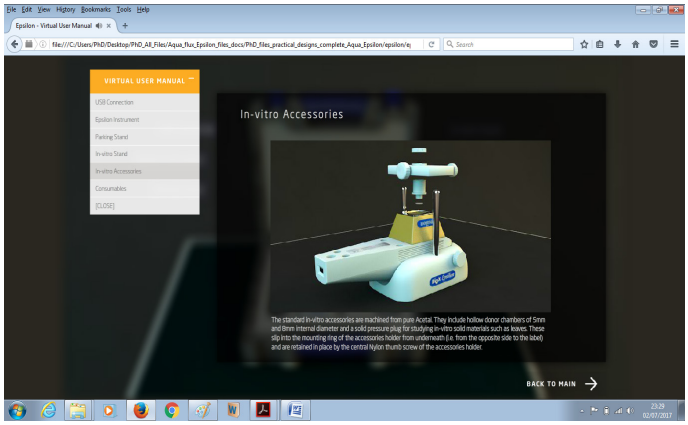


Figure. 44. Epsilon in-vitro stand and accessories demonstrated with the aid of audio feature and 3D ambience in the VUM menu options

Figure 44 shows the VUM of Epsilon in operation from the World Wide Web, with the help of audio and illustrative text displayed in a virtual environment and user interaction, the demonstration process will be very smooth, efficient and comprehensive for all users.

7.3 Epsilon and AquaFlux new holders

The third and final button on the AquaFlux and Epsilon main pages will take the user into the new addition that evolved during the process of working on this project, which is adding a new set of holders to each instrument (AquaFlux and Epsilon) that will certainly enhance the measurement process and will provide a more efficient and accurate reading.

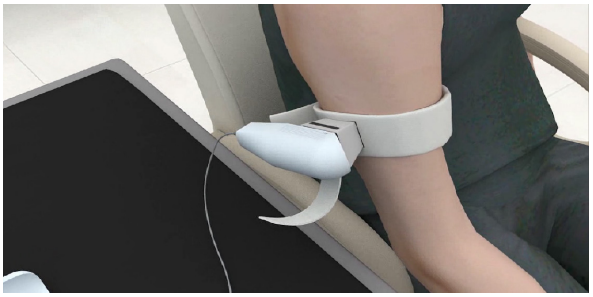


Figure. 45. Epsilon's holder in operation

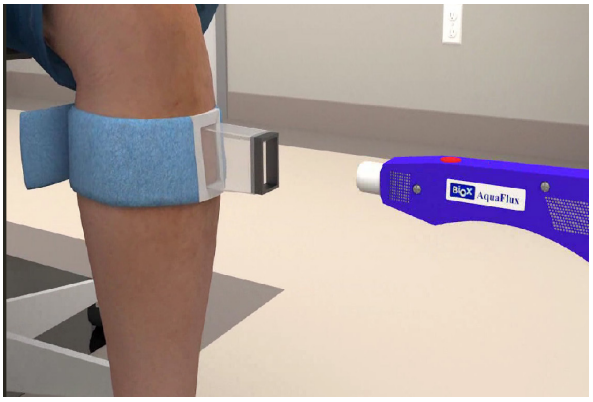


Figure. 46. AquaFlux's holder in operation

7. Evaluating AquaFlux and Epsilon VUM

The VUM of AquaFlux and Epsilon was intended to serve clients, users, and trainees who are interested in purchasing AquaFlux and Epsilon medical instruments and wanted to have a clear and detailed illustrative idea on how these two devices function, and what their features are. From a methodological and marketing point of view, it was vital to conduct a usability study that will show the products' advantages, disadvantages and point out any areas of excellence and parts that require further improvements.

The usability study was conducted by a variety of users, mainly people from a non-IT background. A questionnaire was designed to tackle the most obvious and fundamental questions that could arise while using such a system, moving on to more technical questions. The following table shows the user's feedback on the survey questions that involved 12 independent participants:

	Was the service provided online smooth and error free?	Did you enjoy your experience using the VUM?	Were you successful using the VUM?	Were you able to control the VUM?	Was the audio support feature clear and useful?	Do you think that the VUM will be sufficient to learn how to use the medical devices comprehensively?
1-	Agree	Agree	Neutral	Agree	Agree	Neutral
2-	Neutral	Neutral	Neutral	Neutral	Agree	Neutral
3-	Agree	Agree	Agree	Agree	Agree	Agree
4-	Neutral	Neutral	Neutral	Disagree	Agree	Disagree
5-	Agree	Agree	Agree	Agree	Disagree	Neutral
6-	Agree	Agree	Neutral	Agree	Agree	Neutral
7-	Agree	Neutral	Agree	Agree	Agree	Agree
8-	Agree	Agree	Agree	Agree	Agree	Agree
9-	Agree	Neutral	Agree	Neutral	Agree	Neutral
10-	Agree	Agree	Neutral	Neutral	Agree	Neutral
11-	Agree	Agree	Agree	Neutral	Agree	Disagree
12-	Agree	Neutral	Agree	Agree	Agree	Agree
Agree	10	7	7	7	11	4
Disagree	0	0	0	1	1	2
Neutral	2	5	4	4	0	5

Figure. 47. Survey's questions with user's feedback. 12 users participated

The following illustrative charts show the results of the individual survey questions.

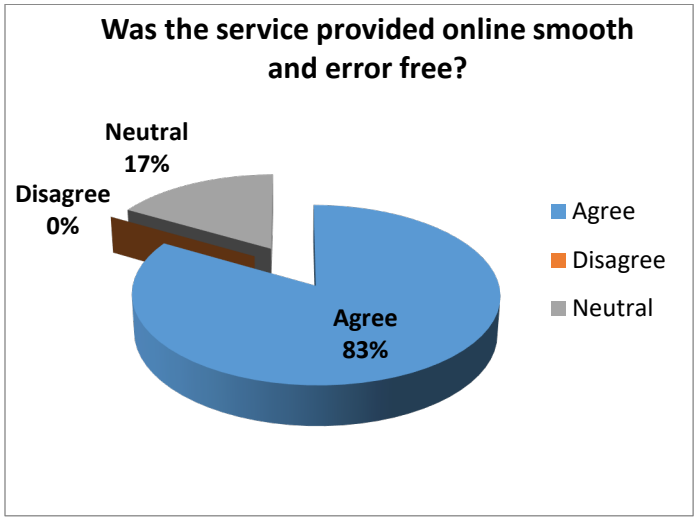


Figure. 48. Shows the success rate of the VUM system being smooth and error-free

The above chart shows that 83% of the participants found that the service provided online was smooth and error-free, whereas 17% were neutral. The result indicates the strength of the virtual system.

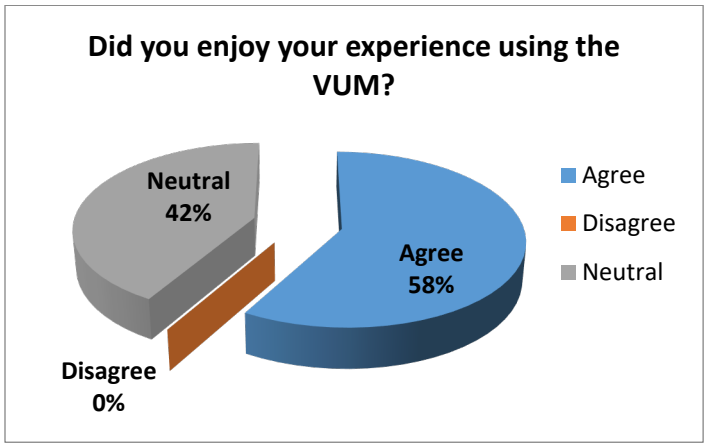


Figure. 49. Shows that the user experience of using VUM was enjoyable

Figure 49 Shows that 58% of the participants enjoyed their VUM experience, 42% were neutral, and 0% disagreed. The result suggests that the system provides an enjoyable experience.

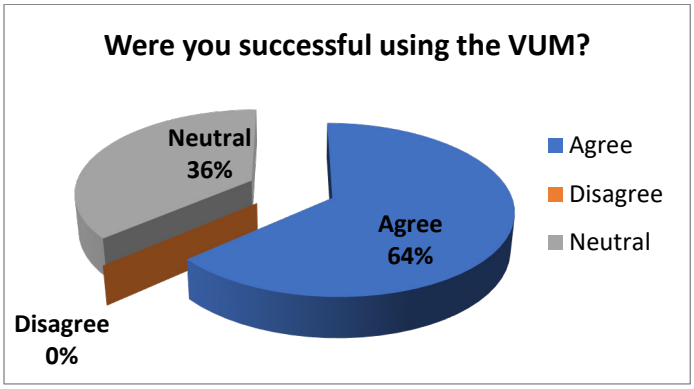


Figure. 50. Shows the success rate of using the VUM

The above chart shows 64% of the participants were successful in using the VUM while 36% were impartial and non disagreed. The chart suggests that the VUM system is a success.

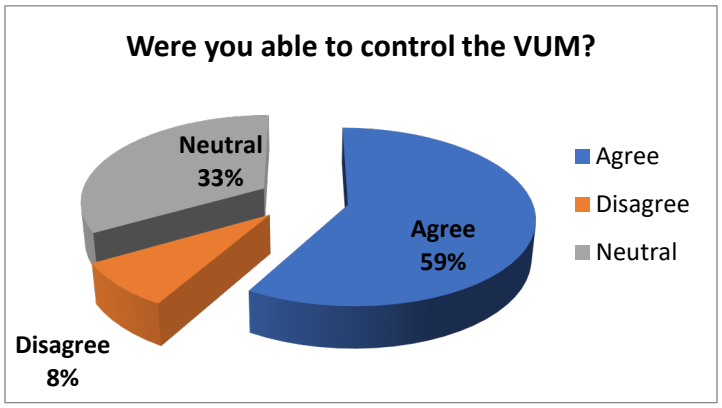


Figure. 51. Shows a high rate of users controlling the VUM

The chart demonstrates that 59% of the participants were able to control the VUM system, whereas 33% were neutral and 8% disagreed. This shows that although a high percentage was able to control it, a minority found it difficult. The result suggests more improvement is required to achieve a higher control success rate.

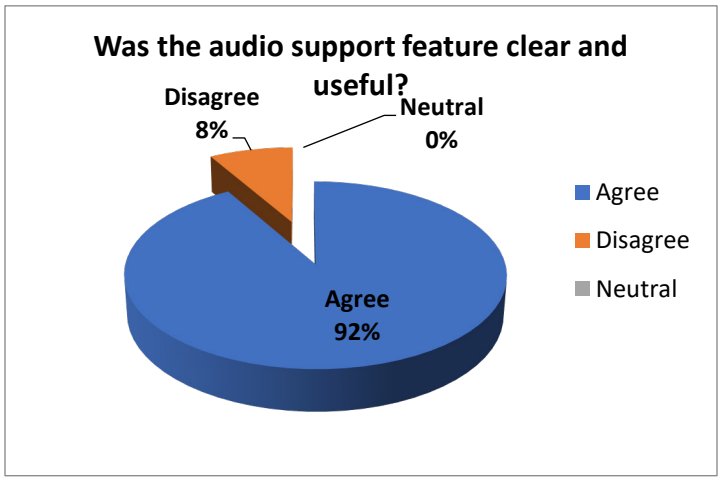


Figure. 52. Shows the success rate of applying the audio feature into VUM

Figure 52 shows that 92% of the participants found that using the support of the audio feature is clear and useful, whereas only 8% disagreed. This is strong evidence to support the excellence of audio feature.

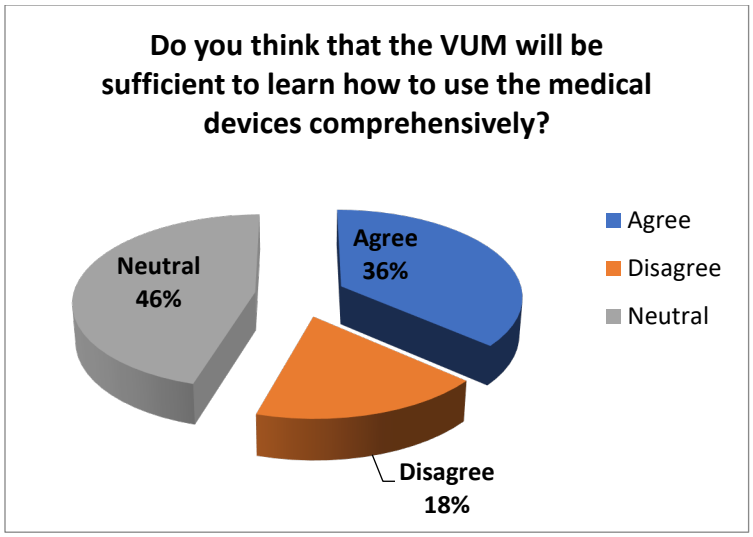


Figure. 53. Shows some users agreed that VUM system is comprehensive

Figure 53 shows that 36% of the participants thought that the VUM system is sufficient and comprehensive, whereas 46% were neutral and 18% disagreed. Although there was a high percentage of neutral responses, a higher percentage agreed rather than disagreed. However, this indicates that there is room for further improvement perhaps by introducing further feature at a later stage.

From the previously illustrated charts and results in relation to all survey's questions, there are two main issues to discuss:

1- Aspects of excellence: presented by
   - The service is smooth and free of errors.
   - Users successfully used the VUM system.
   - The audio feature provided was clear and useful.

2- Aspects of improvements: presented by
   - The VUM is a sufficient tool to teach users how to use the medical devices comprehensively.

The VUM or the interactive virtual environment of both AquaFlux and Epsilon showed great ease of use, efficiency and practicality. With this virtual environment demonstration at hand and its availability online, clients, trainees, and users who are interested in purchasing one of those skin measurement instruments or wanting to see how they operate by obtaining accurate results. In addition, the VUM demonstration process will help immensely in conveying an accurate thought and a precise sense and enjoyable experience of these two devices and how they operate and what their main functions are. Future ideas, improvements, and upgrades are well on the horizon, Augmented Reality (AR), for example, is a fascinating notion that can support this research and enhance it further. Virtual Reality and AR are opening new ways of interactivity between users and the constructed environments [16]. Moreover, Holograms Technology (3D Hologram) provides a remarkable interactivity surface in team works ambience [17].

The results of the survey strongly suggest that the system is operating effectively although there is room for slight improvements.

8. Conclusions

The objective of the research paper was to introduce a total understanding of an online 3D contents and utilising it in an effective way and not limited to displaying it on a webpage. The main contribution to the field is using an interactive virtual environment to demonstrate the process of using a medical device in conducting a human skin measurement for medical purposes with the aid of audio feature. The new web-based 3D interactive VUM for AquaFlux and Epsilon will replace the old method of illustrating how these medical devices work and what are all their features and functionalities via providing each client with an accessible URL after purchasing the instruments. The web-based virtual system is self-explanatory and easy to use as well as the availability of audio support that guides users step by step throughout the process. Novel holders and futuristic concepts have been introduced to display the capabilities of the new addition and to utilise the marketing process using 3D contents with the interactive virtual world to deliver real novel ideas globally.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This research work is part of a postgraduate degree. I would like to take this opportunity to thank London South Bank University for their continuous support and giving us all access needed to use, test, and, experiment these two medical instruments. The authors would like to thank the reviewers for their invaluable comments and recommendations.

References

[1] Al Hashimi, O. and Xiao, P. (2018). Developing a web-based interactive 3D virtual environment for novel skin measurement instruments. *2018 Advances in Science and Engineering Technology International Conferences (ASET),* Abu Dhabi, 2018, pp. 1-8. doi: 10.1109/ICASET.2018.8376823

[2] Cellary, W. and Walczak, K. (2012). *Interactive 3D Multimedia Content*. London: Springer London

[3] VMware Inc, V. (2006). *Virtualization Overview White Paper*. Palo Alto CA: VMware.

[4] Anon, Instrument Development and Digital Signal Processing in Skin Measurements Zhang Xu School of Engineering London South Bank University Supervisors, Oct 2015.

[5] Xiao, P. (2014). Biox Epsilon Model E100: Contact Imaging System. Available at: www.biox.biz/home/brochures.php

[6] Biox (2013). *Epsilon E100Manual*. London: Biox.

[7] Ghanbarzadeh, R., Ghapanchi, A.H. & Blumenstein, M., 2014. Application areas of multi-user virtual environments in the healthcare context. *Studies in Health Technology and Informatics*, 204, pp.38–46.

[8] Shen, W. & Zeng, W., 2011. Research of VR Modeling Technology Based on VRML and 3DSMAX. , pp.487–490.

[9] Neto, C., 2009. Developing Interaction 3D Models for E-Learning Applications.

[10] Kotsilieris, T. & Dimopoulou, N., 2013. " The Evolution of e-Learning in the Context of 3D Virtual Worlds " Department of Health & Welfare Units Administration, Technological Educational. , 11(2), pp.147–167.

[11] Potkonjak, V., Gardner, M., Callaghan, V., Mattila, P., Guetl, C., Petrović, V. and Jovanović, K. (2016). Virtual laboratories for education in science, technology, and engineering: A review. *Computers & Education*, 95, pp.309-327.

[12] Violante, M.G. (2015). Politecnico di Torino Porto Institutional Repository Design and implementation of 3D Web-based interactive.

[13] Naber, J., Krupitzer, C. and Becker, C. (2017). Transferring an Interactive Display Service to the Virtual Reality. Mannheim: IEEE, pp.1-4.

[14] Walczak, K., White, M. & Cellary, W., 2004. Building Virtual and Augmented Reality Museum Exhibitions *. , 1(212), pp.135–145.

[15] Yang, X. et al., 2008. Virtual Reality-Based Robotics Learning System. , (September), pp.859–864.

[16] Voinea, A., Moldoveanu, F. and Moldoveanu, A. (2017). 3D Model Generation of Human Musculoskeletal System Based on Image Processing. In: *21st International Conference on Control Systems and Computer Science*. Bucharest: IEEE, pp.1-3.

[17] Wahab, N., Hasbullah, N., Ramli, S. and Zainuddin, N. (2016). Verification of A Battlefield Visualization Framework in Military Decision Making Using Holograms (3D) and Multi-Touch Technology. In: *2016 International Conference on Information and Communication Technology (ICICTM)*. Kuala Lumpur: IEEE, pp.1-2.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 515-522 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

PID-Type FLC Controller Design and Tuning for Sensorless Speed Control of DC Motor

Abdullah Y. Al-Maliki[*,1,2], Kamran Iqbal[1]

[1]*Department of System Engineering, University of Arkansas-Little Rock, 72204, USA*

[2]*State Company for Oil Project, Iraqi Ministry of Oil, Iraq*

ARTICLE INFO

Article history:
Received: 20 December, 2018
Accepted: 21 December, 2018
Online: 27 December, 2018

Keywords:
DC Motor control
Sensorless speed control
PID
FLC
FLC-PID
Genetic Algorithm
Lookup-Table

ABSTRACT

This article examines the use of non-ideal current and voltage sensors for sensorless speed control for a fixed field DC motor. A PID type speed controller with KF estimator was applied to control the DC motor and IAE, settling time, and peak overshoot were taken as performance indices. However, KF facilitated the noise reduction. After tuning controller gains through MATLAB yielded high peak overshoot as well as IAE with an extended settling time. When we applied, a PID-Type FLC tuned by means of GA (genetic algorithms) caused a 75.98%, 97.89% and 56.2% cut in settling time, maximum overshoot and IAE correspondingly. The FLC-PID fundamentally enhanced sudden load changes disturbance rejection and the reference command speed tracking for the dc motor design in comparison to the conventional PID with no KF. This study was also able to replace the designed FLC-PID with linear lookup-table while achieving the same performance improvements.

1. Introduction

This paper is continuation for our work presented in [1]. DC motors are applied in one way or the other in factories, home appliances, computers to robots, airplanes, and cars. They are more widely used than the related machines, the AC motors, owing to their diverse favorable characteristics. These characteristics some of which are, linear speed control properties and high starting torque. There are more than one types DC motors and all these types have numerous benefits over AC motors which include: less heat production, simpler controllers used, have higher efficiency, can offer precise position control, can produce very close to constant torque and they are easily controllable [2-11]. For that reason, the adoption of DC motors will reduce the amount of energy consumed and improve the efficiency of the machines they are installed. The improvement of DC motors' control arrangement to enhance their response characteristics is one way of achieving these. Is so doing, they will be able to accomplish their work efficiently without the necessarily increasing the capacities of motors alongside their control circuits [12-16].

Researchers have previously done a lot of work to design DC motor control circuits with or without the use of Kalman Filter (KF) and/or Fuzzy Logic Controller (FLC). In [2] the researchers introduced a feedback and feedforward system controllers to control DC motor. While the feedforward loop was just a gain represented by K_f, the feedback loop was PD controller with gains denoted by K_d and K_p. The gains were optimized using fuzzy logic and genetic algorithm and all were constant. A PC was used in [3] to implement a DC motor PID controller, it was configured by trial and error to attain the desired performance. This study also applied a photo sensor to estimate the motor speed and transmit it to the personal computer. In [17] the researchers came up with a torque estimator for DC motors by means of adaptive Kalman filter consisting of two sections; estimation and extraction. Each of these parts is a DC motor's mathematical model and the first one had one more PID controller to guarantee the accuracy of the estimation. Although it required current and speed sensors, this estimator was designed with no torque sensors. A state feedback controller was developed by the researchers in [4] with the use of Extended Kalman Filter (EKF) for induction motor and KF for DC motor to evaluate the states. A DC motor was controlled using a PID controller in [5], K_p K_i K_d were dynamically modified online by three distinct FLC controllers, one for each gain. KF was utilized for tuning the Membership Functions (MFs) of such FLCs. The FLC controllers, in [7,10], were intended to control DC motors. Microcontrollers was used to implement both FLCs and both made use physical sensors ,that is, no estimators were introduced in these studies. In [9], the researchers made use of a Kalman filter, torque sensor and current sensor to measure the

[*]Abdullah Al-Maliki, University of Arkansas-Little Rock, ayalmaliki@ualr.edu

speed of a DC motor and thereafter fed it back to a PI controller to enhance the motor's speed estimation. In [11], the DC motor speed was estimated using Kalman filter and was controlled using both Linear–Quadratic Regulator (LQR) and PID controllers. In [18], the authors adjusted the widths of FLC membership functions by means of arithmetic averaging technique and feed forward to run the DC motor without the use of feedback to minimize online computational needs.

The objective of this study is to realize better transient and steady-state responses for a DC motor's speed control (separately excited brushed type) and efficient reference speed tracking irrespective of load and alteration of reference speed. The voltage will only be used as the output of the controller, while the current will not be controlled but restricted within the rated motor current. Then the results from the PID controller and conventional PID controller will be compared with the PID-type FLC controller (FLC-PID) and KF. In addition to the classical time response characteristics; the performance index of IAE is selected to illustrate the achievement in the motor speed response. IAE displays the total error for the entire run period of the simulation to increase the size of the built controller in tracking command speed, notwithstanding its variations and changes in load. In Section 2 below, the DC motor model is developed, Section 3 covers the KF design, Section 4 addresses the different controller designs, and Section 5 carries the conclusion.

2. DC Motor Model

The most common type of DC motors which are broadly applied in robotic and industrial functions is the separately excited DC motor with constant excitation field can be represented by state-space model [4,13]. This model comprises two differential equations representing the mechanical and electrical responses of the DC motor in both static and active operating states. Equations (1) and (2) illustrate this model [4,19].

$$\begin{bmatrix} \dot{I}_a \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} -\frac{R_a}{L_a} & -\frac{K_b}{L_a} \\ \frac{K_t}{J_r} & -\frac{B_m}{J_r} \end{bmatrix} \cdot \begin{bmatrix} I_a \\ \omega \end{bmatrix} + \begin{bmatrix} \frac{1}{L_a} & 0 \\ 0 & -\frac{1}{J_r} \end{bmatrix} \cdot \begin{bmatrix} V \\ T_{load} \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} I_a \\ \omega \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} I_a \\ \omega \end{bmatrix} \quad (2)$$

The chosen motor specification are listed in Table 1 [20].

The values from Table I are substituted in (1) and (2) and a sampling time (T_s) of 1 millisecond was chosen. Then, using a MATLAB (c2d) function, the continues state-space equations are changed to discrete. The discrete-time state-space motor model is shown in (3) and (4):

$$\begin{bmatrix} i_a(t+1) \\ \omega(t+1) \end{bmatrix} = \begin{bmatrix} 0.9526 & -0.0231 \\ 0.0153 & 0.9998 \end{bmatrix} \cdot \begin{bmatrix} i_a(t) \\ \omega(t) \end{bmatrix} + \begin{bmatrix} 0.0210 & 0.0002 \\ 0.0002 & -0.0143 \end{bmatrix} \cdot \begin{bmatrix} v(t) \\ t_{load}(t) \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} i_a(t) \\ \omega(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} i_a(t) \\ \omega(t) \end{bmatrix} \quad (4)$$

Where (t) represents the current sample, and (t+1) denote the next sample.

Table 1 Chosen motor specifications [20]

Parameter/Specification	Value	Unit
R_a (Armature resistance)	2.25	Ohm
L_a (Armature inductance)	$46.5 \cdot 10^{-3}$	H
K_b (Back EMF constant)	1.1	Volt · sec/rad
K_t (Torque constant)	1.1	N · m/A
J_r (Rotor inertia)	$7 \cdot 10^{-2}$	Kg · m^2
B_m (Mechanical damping factor)	$2 \cdot 10^{-3}$	N · m · sec/rad
V_R (Rated Voltage)	220	V
T_R (Rated Torque)	4.7495	N · m
I_R (Rated current)	4.8	A
ω_R (Rated angular speed)	157.08	rad/sec
P_R (Rated power)	1	HP

3. Kalman Filter Observer

As mentioned in [21,22] Kalman filter is considered as optimum observer for the state variables and the outputs of linear time-invariant (LTI) systems. It has two significant characteristics which made it our choice to estimate the states of the DC motor under study; one it is able to exploit the already known state-space model of the DC motor, and the second reason because it has the capacity to filter diverse sensor noises [23]. The Kalman filter system discrete equations can be broken into two categories and are illustrated in (5)-(10) [23].

A. Time update equations

$$\hat{x}_t = A \cdot x_{t-1} + B \cdot u_t + W_t \quad (5)$$

$$\hat{P}_t = A \cdot P_{t-1} \cdot A^T + Q_t \quad (6)$$

B. Measurement update equations

$$K_t = \frac{\hat{P}_t \cdot H^T}{H \cdot \hat{P}_t \cdot H^T + R} \quad (7)$$

$$y_t = C_{KF} \cdot x_{m_t} + z_k \quad (8)$$

$$x_t = \hat{x}_t + K_t \cdot (y_t - H \cdot \hat{x}_t) \quad (9)$$

$$P_t = (I - K_t \cdot H) \cdot \hat{P}_t \quad (10)$$

In which $\hat{x}_t$ is the estimated current state vector; $\hat{P}_t$ is the estimated process error covariance matrix; W_t is the estimated noise matrix; Q_t is the system noise covariance matrix; K_t is Kalman gain; H: is the transformation matrix; R is sensor noise covariance matrix; y_t is measurement of the state; C_{KF} is Kalman filter C matrix for state-space model of the system under study (in this case the DC motor); x_{m_t} is the measured states vector; z_k is measurement noise; x_t is corrected state vector; and, P_t is corrected process error covariance matrix

Kalman filter estimates the states of present system and the error in such estimates by means of the time update equations. It then rectifies these errors using the measurement update calculations based on the Kalman gain and the actual inaccurate measurement. This done in a repetitive way for all sample times (T_s) or time steps [23]. Let $\Delta\omega$ and ΔI_a represent the error in speed and current calculation/measurement correspondingly.

Then accurate T_{load} computation needs the integration of ω and $d\omega/dt$ as illustrated in (11), which is solving the second row of (1) for T_{load}:

$$T_{load} = -\frac{d\,\omega}{dt} J_r + K_t \cdot I_a - B_m \cdot \omega \quad (11)$$

As $|K_t \cdot I_a| \gg |-B_m \cdot \omega|$ and $|K_t \cdot I_a| \gg \left|-\frac{d\,\omega}{dt} J_r\right|$, and the term $-\frac{d\,\omega}{dt} J_r$ has noise because of the differentiation, The approximate formula below can be used to solve T_{load}:

$$T_{load} \cong K_t \cdot I_a \quad (12)$$

Equation (13) is derived by solving the first row of (3) for ω, is used to calculate the motor speed:

$$\omega = -\frac{L_a}{K_b}\frac{d\,I_a}{dt} - \frac{R_a}{K_b} I_a + \frac{1}{K_b} V \quad (13)$$

Because $\left|-\frac{R_a}{K_b} I_a + \frac{1}{K_b} V\right| \gg \left|-\frac{L_a}{K_b}\frac{d\,I_a}{dt}\right|$, ω can be calculated with high enough accuracy using (14) below:

$$\omega \cong -\frac{R_a}{K_b} I_a + \frac{1}{K_b} V \quad (14)$$

The sensors and process noise covariance matrices are articulated in terms of the stochastic characteristics of their respective noises and they can be either static or dynamic [24]. Q elements were found to be directly proportional to the inverse of the estimated armature current and motor speed and therefore was chosen to be dynamic as shown in (15). Since the sensor error is fixed by the manufacturer in its datasheet, R was selected to be static. In (16) the calculations of R are shown and the datasheet of a sample current sensor is found in [25].

$$Q = \begin{bmatrix} \frac{1}{I_a^2} & \frac{1}{I_a^2}\cdot\frac{1}{\omega^2} \\ \frac{1}{I_a^2}\cdot\frac{1}{\omega^2} & \frac{1}{\omega^2} \end{bmatrix} \quad (15)$$

$$R = \begin{bmatrix} 2.25 & 0 \\ 0 & 6.25 \end{bmatrix} \quad (16)$$

4. Controller Design, Simulation and Results

In this section, three potential designs of a controller for DC motor speed control are discussed. Figure 1 shows the general system architecture under study.

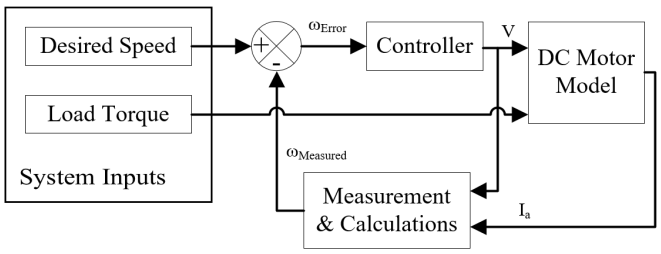


Figure 1: The DC motor with the PID controller and speed feedback

4.1. Discrete PID Controller

Initially, discrete PID was used to control the DC motor. The transfer function of the controller is presented in (17). Figure 2 shows the command speed signal to the motor, the motor will experience an uncontrolled torque load variation illustrated in Figure 3. In the rest of this work, Figure 2 and Figure 3 will be considered as the running conditions for the DC motor under study.

Using MATLAB/Simulink PID tuner with "design focus" option set to "Reference tracking", a discrete PID controller was tuned to control the motor's speed by means of changing its supply voltage. The gains after MATLAB/Simulink PID tuner are: K_P=2.51, K_I=9.724, K_D=-0.19185 and N=12.89.

$$U(Z) = K_P + K_I \cdot T_s \cdot \left(\frac{1}{z-1}\right) + K_D \cdot \left(\frac{N}{1+N\cdot T_s\cdot\left(\frac{1}{z-1}\right)}\right) \quad (17)$$

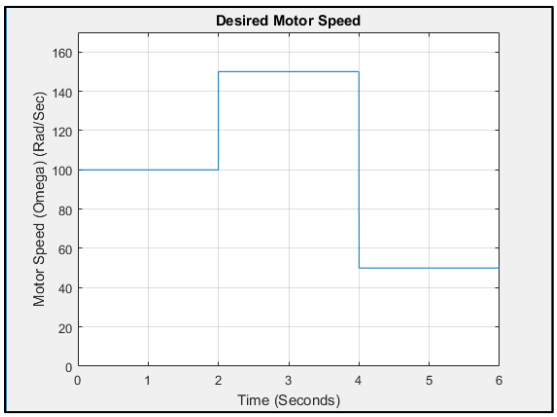


Figure 2 Desired motor speed profile

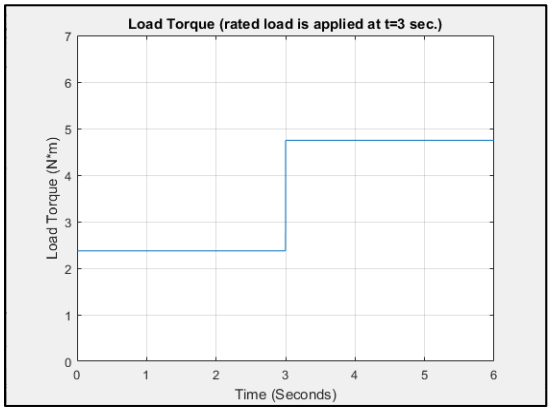


Figure 3 Load torque (disturbance) profile

To simulate current and voltage sensors precision tolerances, arbitrary noise of 2% was added to voltage and current [25]. The voltage (controller output) was kept at or below the nameplate voltage that is 220 V, and this limitation is kept for the FLC PID in part C and throughout this work. The measurement of current was restricted certain upper limit to mimic the current sensor magnetic limitation characteristics. In the measurement block, the speed is determined using (13) only and no more conditioning is required. Figure 4 demonstrate the system response in which the system experienced an IAE of 47.52, settling time (t_s) of 1.08 second and peak overshoot of 25.7%. Using (14) to obtain ω in the subsequent run lowered the peak overshoot to 23.7% and lowered the IAE a little making it 45.89. The settling time experienced a slight fall to 1.07 seconds. Such a small enhancement in system response is associated to the cancelation of the noisy derivative section in (13). The measured and actual motor speed using (13) and with no Kalman filter are displayed in Figure 5.

4.2. PID with KF

Kalman filter was applied in the next run in an effort to predict the motor's speed, Figure 6 shows the whole system. When Kalman filter was used and everything else as from the previous run was unchanged, Figure 7 displays clear falls in the settling time, IAE and peak overshoot which are now in order, 0.635, 36.71

and 8.01%. The KF estimated speed of the motor vs the actual and measured speeds are shown in Figure 8.

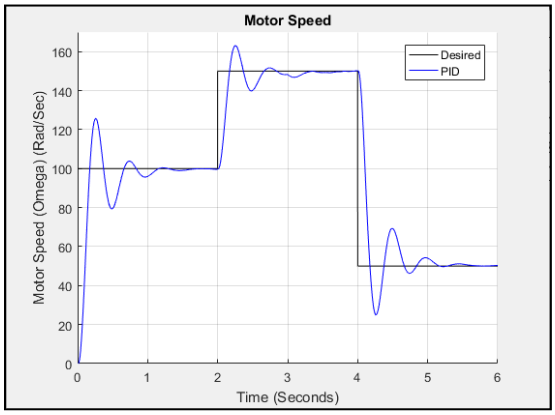


Figure 4 motor speed response with PID and unit feedback

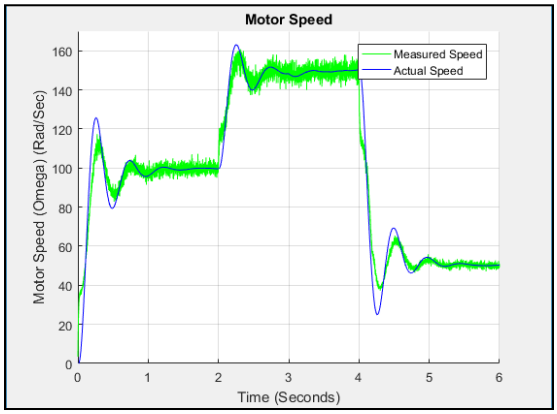


Figure 5 actual motor speed and measured motor speed with added noise

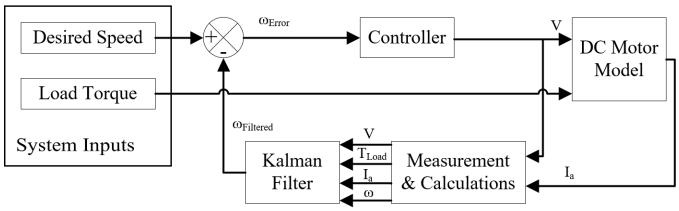


Figure 6 The DC motor with the PID controller and Kalman filter

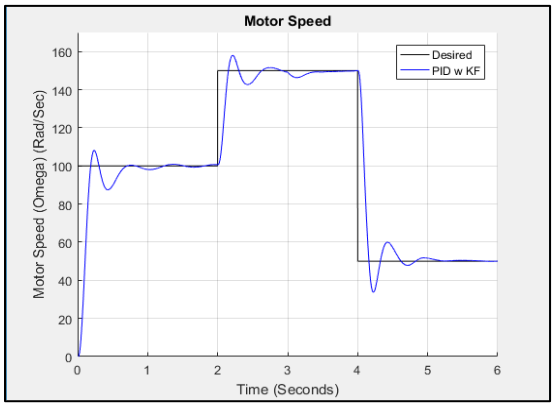


Figure 7 motor speed response with PID and Kalman filter estimator

While it showed some improvement in the system response, the PID with the help of Kalman filter was unable to fully remove the overshoot of the motor's speed response which introduced unwanted ringing in the motor's speed. This limitation can be overcome by designing a PID-type fuzzy logic controller with PID properties and the additional variable gains throughout the full range of the controller's operation [26,27].

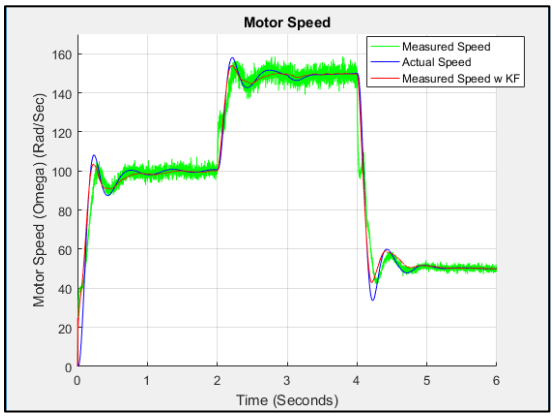


Figure 8 actual motor speed and measured motor speed with added noise with and without Kalman filter

4.3. FLC-PID Controller

A Takagi-Sugeno type FLC-PID controller is chosen in this work as it is "more diverse in gain variation characteristics" than the Mamdani type FLC-PID controller [26]. Aside from few changes in the rule base; the main difference in design between Takagi-Sugeno type FLC-PID controller developed in this work and the one utilized in [28,29] is the different and distinct gains for PI and PD controllers, offering an additional degree of design freedom. Figure 9 represents the block diagram of the FLC-PID developed for this work.

For both controllers (FLC-PI and FLC-PD), the input and output signal ranges were normalized to [-1,1]. Doing so will enable this FLC-PID controller to be applied for various motor sizes by just tuning its gains. Figure 10 shows the MF's for error as well as rate of change of error, Gaussian type was chosen to guarantee smooth control action. The MF's for the controller's output are presented in Figure 11 which are uni-valued MF's. one benefit of using FLC controller is that we can make all the equivalent MF's similar for both FLC-PD and FLC-PI, and the unique action variation between the controllers is achieved only by altering their rule base. Which is what we did in this work.

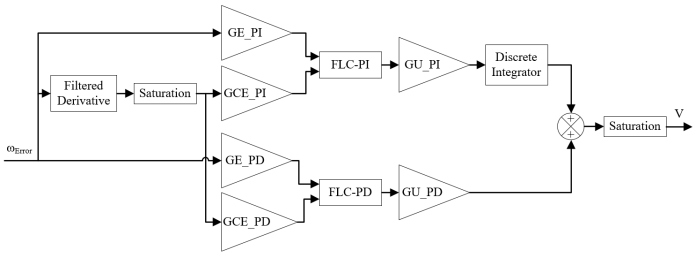


Figure 9 FLC-PID block diagram

Where GE_PI is FLC-PD speed error gain; GCE_PI is FLC-PD change of rate of speed error gain; GU_PI is FLC-PD output gain; GE_PD is FLC-PD speed error gain; GCE_PD is FLC-PD change of rate of speed error gain; and, GU_PD is FLC-PD output gain.

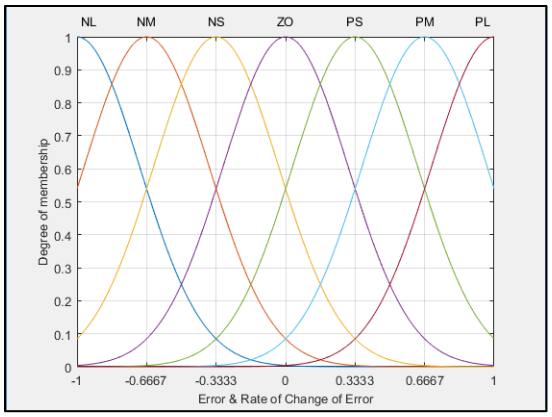


Figure 10 input membership functions for FLC-PI and FLC-PD

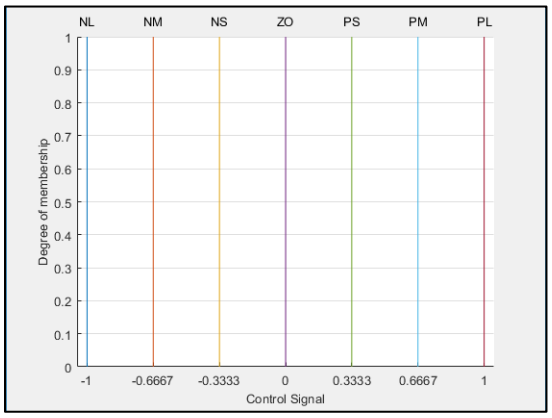


Figure 11 output membership functions for FLC-PI and FLC-PD

The rule base for FLC-PD and FLC-PI are shown by Tables 2 and 3 respectively. The FLC-PID applies these rules in the arrangement illustrated in (17) [30]to produce the control surfaces presented in Figure 12 and 13. Such rules simulates the performance of the common PI and PD controllers to develop FLC-PI and FLC-PD controllers in that order.

Table 2 FLC-PD rule base

Control Output UPD		Error (e)						
		NL	***NM***	***NS***	***ZO***	***PS***	***PM***	***PL***
Rate of Change of Error (ce)	***PL***	PS	PS	PM	PL	PL	PL	PL
	PM	PS	PS	PS	PS	PL	PL	PL
	PS	PM	PS	PS	PS	PM	PL	PL
	ZO	NL	NM	PS	ZO	PS	PM	PL
	NS	NL	NL	NM	NS	PS	PS	PM
	NM	NL	NL	NL	NS	PS	PS	PS
	NL	NL	NL	NL	NS	PS	PS	PS

Table 3 FLC-PI rule base

Control Output UPI		Error (e)						
		NL	***NM***	***NS***	***ZO***	***PS***	***PM***	***PL***
Rate of Change of Error (ce)	***PL***	ZO	PS	PL	PL	PL	PL	PL
	PM	NS	ZO	PS	PM	PL	PL	PL
	PS	NL	NS	ZO	PS	PM	PL	PL
	ZO	NL	NM	NS	ZO	PS	PM	PL
	NS	NL	NL	NM	NS	ZO	PS	PL
	NM	NL	NL	NL	NM	NS	ZO	PS
	NL	NL	NL	NL	NL	NL	NS	ZO

Where PL is Positive Large; PM is Positive Medium; PS is Positive Small; ZO is Zero; NS is Negative Small; NM is Negative Medium; and NL is Negative Large.

$$\underbrace{\text{if } \underbrace{\text{Error}}_{\substack{\text{1st}\\\text{input}}} = \underbrace{\text{NL}}_{\substack{\text{1st}\\\text{input}\\\text{MF}}} \underbrace{\text{AND}}_{\substack{\text{Fuzzy}\\\text{Logic}\\\text{Operator}}} \underbrace{\text{Rate of Change of Error}}_{\substack{\text{2nd}\\\text{input}}} = \underbrace{\text{PL}}_{\substack{\text{2nd}\\\text{input}\\\text{MF}}}}_{\text{Antecedent}} \text{ Then } \underbrace{\underbrace{\text{UPD}}_{\text{Output}} = \underbrace{\text{PS}}_{\substack{\text{Output}\\\text{MF}}}}_{\text{Consequent}} \quad (17)$$

When we have (17), Tables 2 and 3, Figure 10, 11, 12 and 13 into perspective; it would be obvious that both FLC have more than 49 equations and more than 10 parameters (comprising membership functions parameters) to adapt its output. These add to diverse design flexibility for the preferred control surface that the common PID controller simply cannot provide.

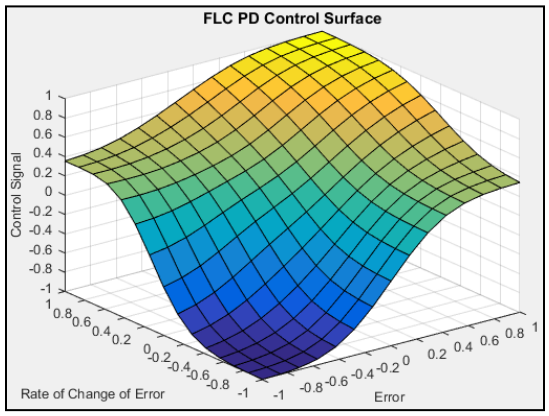


Figure 12 FLC-PD control surface

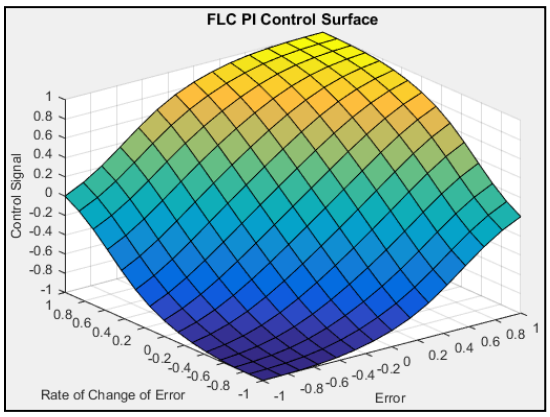


Figure 13 FLC-PI control surface

The next five steps briefly describe how an FLC works [30,31,32]:

1. The input signals are mapped to their corresponding MF's and assign degree of membership ranging between [0,1] for each input signal. (Fuzzification stage)
2. Finding the degree of firing of each rule by applying fuzzy logic operation (e.g. AND, OR, … etc.) to the antecedents. (Inference mechanism matching stage1).
3. FLC determines which rule is to be fired by checking the result from step 2, each rule having a nonzero result will be fired. (Inference mechanism matching stage 2).
4. Multiply the consequent of each rule by its firing degree (the corresponding result from step 3). (Inference setup/aggregation stage).

5. Apply the selected defuzzification method (for this work weighted average method was selected, which is shown in equation 18). (Defuzzification stage).

$$u = \frac{\sum_{n=1}^{\text{no.of fired rules}} \mu_{\text{rule}_n} \cdot U_{\text{rule}_n}}{\sum_{n=1}^{\text{no.of fired rules}} \mu_{\text{rule}_n}} \quad (18)$$

where:
u: is FLC output
μ_{rule_n}: is the firing degree for each rule (step 2 above)
U_{rule_n}: is the consequent of rule_n.

The selection of FLC-PID gains was informed by the following: because the rated speed of the motor is 157.08 rad/sec, speed error gains were calculated as $\frac{1}{157.08} = 0.006367$. A range of [-1000,1000] was taken for derivative of error to neutralize high values due to the derivative term; consequently, it's gain was chosen to be $\frac{1}{1000} = 0.001$. Lastly, the output gains is set to be $1.5 * \text{rated voltage}$ to put an accelerated control action to the motor's speed. FLC-PID gains first estimation may as well be acquired from an adjusted PID as stated in [28].

4.4. Gains' Value Optimization Using GA

Genetic algorithm (GA) is an evolutionary algorithm; it uses the process of natural selection to solve optimization problems; it works to find the fittest value in the range of search regardless of that range's nonlinearity. Simple GA can be summarized in the following steps [33]:

1. Convert the variable values into binary. (Initialization 1)
2. Initiate random population of 50 individual per variable. (Initialization 2)
3. Check the fitness of each individual. (Fitness 1)
4. Select the best-fitted individuals. (Fitness 2)
5. Generate new individuals (children) by combining two individuals (parents) from step 4. (Crossover)
6. Generate new individuals (children) by making random changes to individuals (parents) from step 4. (Mutation)
7. Replace the current population with the children from steps 4 and 5. (Initiate new generation)
8. Repeat from step 3 until at least one stopping condition (number of generations, the value of the relative change in the fitness function, … etc.) is satisfied.

Table 4 tuned controller gains

Gain	Value
GE_PI	0.1
GCE_PI	$8.51 \cdot 10^{-4}$
GU_PI	225
GE_PD	$2.43 \cdot 10^{-2}$
GCE_PD	$1.134 \cdot 10^{-3}$
GU_PD	300

MATLAB optimization tool with built-in GA optimizer has been used in this project to fine tune the FLC-PID gains. The selected fitness function was the IAE of the speed, to ensure that the tuning is aimed at faster and improved speed command tracking with an all-round load change rejection. Because GA is a stochastic search tool, various runs may produce varying results, but each result must satisfy a minimum of one stopping condition. Indeed, the remaining runs with various gains can result in similar system response. Table 4 shows the FLC-PID gains upon tuning.

The FLC-PID successfully removed the peak overshoot (less than 0.5%), decreased the settling time to 0.257 seconds and reduced the IAE to only 20.1. To make this clearer, Table 4 illustrates a percentage contrast between FLC-PID with KF, common PID without KF and PID with KF, choosing common PID without KF as a 100% base. Figure 14 is a graph of the motor speed response for all scenarios mentioned above versus the desired speed. It is worth noting that lower value for every performance index used in this paper means improved system response.

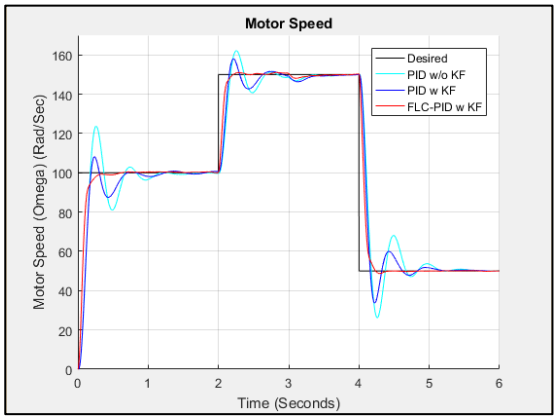


Figure 14 compare motor speed for all three controllers used in this work

Table 4 comparison of system response for all three controllers used in this work

Performance Indices	PID W/O KF	PID With KF	FLC-PID With KF
Settling time (seconds)	1.07 100%	0.635 59.35%	0.257 24.02%
Peak Overshoot (radians)	23.7 100%	8.01 33.8%	<0.5 <2.11%
IAE	45.89 100%	36.71 80%	20.1 43.8%

4.5. Converting FLC-PID to lookup Table

In order to reduce the computational requirement of FCL-PID to a real word controller e.g. microcontroller or field programmable gate array (FPGA), thus making it easier for the controller to response in real time, the FLC-PI and FLC-PID control surfaces will be stored in a linear lookup-Table format. This can be done by dividing the two inputs of both FLC controllers (FLC-PI and FLC-PD) into 21 breakpoints (since both inputs range is normalized to $[-1,1]$, 21 breakpoints mean 0.1 spacing) and evaluate the FLC controllers at each of the breakpoints. The total evaluations will be $21 * 21 = 441$ for each FLC controller. A linear interpolation based lookup Table was put in place of each FLC controller in the controller block as shown in Figure 15, and they were loaded with the corresponding 441 evaluations. The motor speed response with FLC-PID and lookup Table FLC-PID are shown in Figure 16. Both responses are actually overlapping as the difference between them is minimal. A zoomed view of the speed response from the time 0.3to 0.7 seconds is shown in Figure 17 to magnify the difference between these responses, which has a maximum value of less than 1%. As a

result, the DC motor with the lookup Table had almost identical performance indices (the difference is less than 0.1%) as compared to the response with the actual FLC-PID controller.

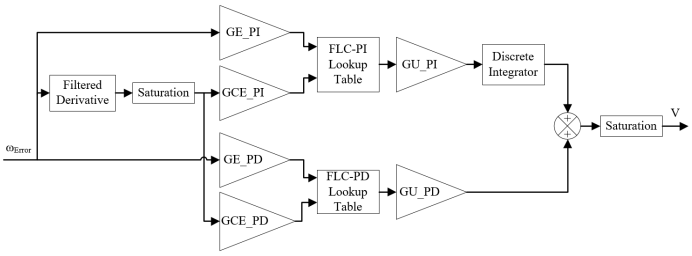


Figure 15 lookup table FLC-PID block diagram

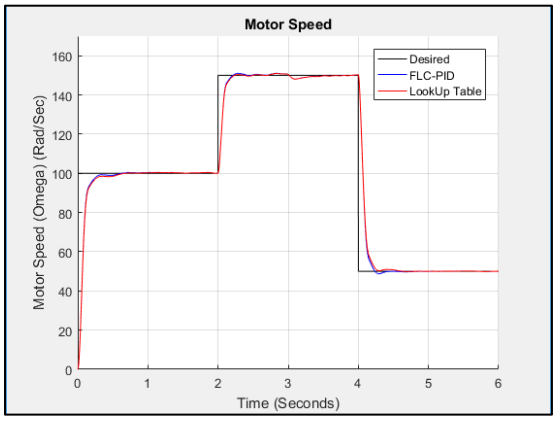


Figure 16 lookup table FLC-PID versus actual FLC-PID speed response

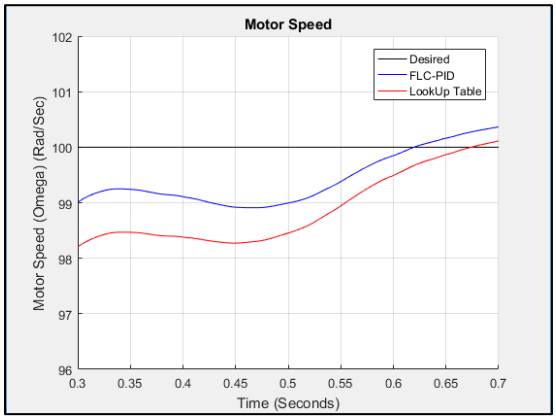


Figure 17 zoomed view of lookup Table FLC-PID versus actual FLC-PID speed response

5. Conclusion

This work investigated the sensorless speed control of fixed field DC motor in a noisy environment mimicking real-live operating environment. As a result of noise in this environment, a common PID controller yielded high peak overshoot from the reference and took significate time to settle. That was true even after removing the noisy derivative terms in T_{load} and ω in their respective estimation equations. When using Kalman filter, the PID controller was able to lower the peak overshoot by good margin, nonetheless it stayed there with long ringing in the speed response. To minimize these undesirable system behaviors , an FLC-PID controller was applied to the motor. This controller was tuned by GA. This new controller design end up reducing the peak overshoot to less than 1% and the settling time by 75.98%. The IAE was also reduced to only 56.2% of what it was. Finally, the FLC-PID was converted into linear interpolation- based lookup-Table to reduce the computational complexity of the FLC-PID and make it easier to realize in real world applications while maintaining the same response as with the actual FLC-PID. Achieving these enhanced time response and speed tracking improvement will ensure smoother, more robust and more power saving operation of the motor. The FLC-PID gave the DC motor "immunity" to reference speed and load changes as long as these changes are within its operating capacity.

In the future we intend to implement the lookup-Table by using a specialized lookup-Table Integrated Circuit (IC) or a generic Field Programmable Gate Array (FPGA) IC and to connect it to a physical separately excited DC motor in an experimental setup. Variable load, variable reference speed and current limiting protection must be included in this setup. This implementation will require retuning the FLC-PID gains, but in return it will solidly verify the proposed controller.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Y. Al-Maliki and K. Iqbal, "FLC-based PID controller tuning for sensorless speed control of DC motor," Proc. IEEE Int. Conf. Ind. Technol., vol. 2018–February, pp. 169–174, 2018.
[2] S. Chowdhuri and A. Mukherjee, "An evolutionary approach to optimize speed controller of DC machines," Proc. IEEE Int. Conf. Ind. Technol. 2000 (IEEE Cat. No.00TH8482), vol. 2, pp. 682–687, 2000.
[3] G. Huang and S. Lee, "PC-based PID speed control in DC motor," 2008 Int. Conf. Audio, Lang. Image Process., pp. 400–407, 2008.
[4] G. G. and P. Siano, "Sensorless Control of Electric Motors with Kalman Filters: Applications to Robotic and Industrial Systems," Int. J. Adv. Robot. Syst., vol. 8, no. 6, p. 1, 2011.
[5] M. Shadkam, H. Mojallali, and Y. Bostani, "Speed Control of DC Motor Using Extended Kalman FilterBased Fuzzy PID," Int. J. Inf. Electron. Eng., vol. 3, no. 9, pp. 1209–1220, 2013.
[6] J. Jacob, S. S. Alex, and A. E. Daniel, "Speed Control of Brushless DC Motor Implementing Extended Kalman Filter," Int. J. Eng. Innov. Technol., vol. 3, no. 1, pp. 305–309, 2013.
[7] A. A. Thorat, S. Yadav, and S. S. Patil, "Implementation of Fuzzy Logic System for DC Motor Speed Control using Microcontroller," J. Eng. Res. Appl., vol. 3, no. 2, pp. 950–956, 2013.
[8] J. N. Rai and M. Singhal, "Speed Control of Dc Motor Using Fuzzy Logic Technique," IOSR J. Electr. Electron. Eng., vol. 3, no. 6, pp. 41–48, 2012.
[9] P. Deshpande and A. Deshpande, "Inferential control of DC motor using Kalman Filter," in 2012 2nd International Conference on Power, Control and Embedded Systems, 2012, pp. 1–5.
[10] H. R. Jayetileke, W. R. De Mel, and H. U. W. Ratnayake, "Real-time fuzzy logic speed tracking controller for a DC motor using Arduino Due," 2014 7th Int. Conf. Inf. Autom. Sustain. "Sharpening Futur. with Sustain. Technol. ICIAfS 2014, 2014.
[11] T. Abut, "Modeling and Optimal Control of a DC Motor," Int. J. Eng. Trends Technol., vol. 32, no. 3, pp. 146–150, Feb. 2016.
[12] A. O. Al-Jazaeri, L. Samaranayake, S. Longo, and D. J. Auger, "Fuzzy Logic Control for energy saving in Autonomous Electric Vehicles," in 2014 IEEE International Electric Vehicle Conference (IEVC), 2014, pp. 1–6.
[13] A. Watanabe, S. Yuta, and A. Ohya, "A new method for efficient drive and current control of small-sized brushless DC motor: Experiments and its evaluation," in IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, 2010, pp. 735–741.
[14] L. Varghese and J. T. Kuncheria, "Modelling and design of cost efficient novel digital controller for brushless DC motor drive," in 2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD), 2014, pp. 1–5.
[15] Y.-P. Yang and M.-T. Peng, "A Surface-Mounted Permanent-Magnet Motor With Sinusoidal Pulsewidth-Modulation-Shaped Magnets," IEEE Trans.

Magn., vol. PP, pp. 1–8, 2018.
[16] S. Noguchi and H. Dohmeki, "Improvement of efficiency and vibration noise characteristics depending on excitation waveform of a brushless DC motor," 2018 IEEE Int. Magn. Conf., pp. 1–5, 2018.
[17] S. Lee and H. Ahn, "Sensorless torque estimation using adaptive Kalman filter and disturbance estimator," in Proceedings of 2010 IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications, 2010, pp. 87–92.
[18] J. K. Satapathy, S. Das, C. J. Harris, and P. Misra, "Indirect tuning of membership function in a fixed fuzzy structure for efficient control of a DC drive system," in SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. "Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions" (Cat. No.00CH37166), 2000, vol. 5, pp. 3790–3793.
[19] S. P. Paul C. Krause, Oleg Wasynczuk, Scott D. Sudhoff, Analysis of Electric Machinery and Drive Systems, 3rd ed. Piscataway, NJ, USA: John Wiley & Sons, Inc., 2013.
[20] F. Rashidi, M. Rashidi, and A. Hashemi-Hosseini, "Speed regulation of DC motors using intelligent controllers," Control Appl., pp. 925–930, 2003.
[21] F. L. Lewis, L. Xie, and D. Popa, Optimal and Robust Estimation With an Introduction to Stochastic Control Theory, Second Edi. CRC press, 2008.
[22] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," J. Basic Eng., vol. 82, no. 1, p. 35, 1960.
[23] R. G. Brown and P. Y. C. Hwang, Introduction to Random Signals and Applied Kalman Filtering WITH MATLAB EXERCISES, FOURTH EDI. John Wiley & Sons, Inc., 2012.
[24] Texas Instruments and T. I. Europe, "Sensorless Control with Kalman Filter on TMS320 Fixed-Point DSP," Control, no. July, 1997.
[25] "Split Core Current Transformer ECS1030-L72." ECHUN Electronic Co., Ltd, Dongguan City. Guang Dong. China, pp. 1–3, 2017.
[26] H. Ying, "Constructing nonlinear variable gain controllers via the Takagi-Sugeno fuzzy control," IEEE Trans. Fuzzy Syst., vol. 6, no. 2, pp. 226–234, 1998.
[27] B. M. Mohan and A. Sinha, "Analytical structure and stability analysis of a fuzzy PID controller," Appl. Soft Comput., vol. 8, no. 1, pp. 749–758, 2008.
[28] A. Noshadi, J. Shi, W. Lee, and A. Kalam, "PID-type fuzzy logic controller for active magnetic bearing system," in IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society, 2014, vol. 27, no. 7, pp. 241–247.
[29] J. X. Xu, C. C. Hang, and C. Liu, "Parallel structure and tuning of a fuzzy PID controller," Automatica, vol. 36, no. 5, pp. 673–684, 2000.
[30] K. M. Passino and S. Yurkovich, Fuzzy Control. Addison Wesley Longman, Inc., 1998.
[31] G. S. Sandhu, T. Brehm, and K. S. Rattan, "Analysis and design of a proportional plus derivative fuzzy logic controller," in Proceedings of the IEEE 1996 National Aerospace and Electronics Conference NAECON 1996, 1996, vol. 1, pp. 397–404.
[32] S. Z. Hassan, H. Li, T. Kamal, F. Mehmood, "Fuzzy embedded MPPT modeling and control of PV system in a hybrid power system" in 12th International IEEE Conference on Emerging Technologies (ICET) 18-19-October 2016
[33] D. A. Coley, An Introduction to Genetic Algorithms for Scientists and Engineers. WORLD SCIENTIFIC, 1999.