Lecture Notes in Electrical Engineering 481

Rayner Alfred
Yuto Lim
Ag Asri Ag Ibrahim
Patricia Anthony *Editors*

Computational Science and Technology

5th ICCST 2018, Kota Kinabalu, Malaysia, 29–30 August 2018

Springer

Lecture Notes in Electrical Engineering

Volume 481

** **Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, MetaPress, Springerlink** **

Lecture Notes in Electrical Engineering (LNEE) is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering
- Engineering

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer's other Lecture Notes series, LNEE will be distributed through Springer's print and electronic publishing channels.

For general information about this series, comments or suggestions, please use the contact address under "service for this series".

To submit a proposal or request further information, please contact the appropriate Springer Publishing Editors:

More information about this series at http://www.springer.com/series/7818

Rayner Alfred · Yuto Lim
Ag Asri Ag Ibrahim · Patricia Anthony
Editors

Computational Science and Technology

5th ICCST 2018, Kota Kinabalu, Malaysia, 29–30 August 2018

Springer

Editors
Rayner Alfred
Knowledge Technology Research Unit,
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu, Sabah, Malaysia

Yuto Lim
School of Information Science,
Security and Networks Area
Japan Advanced Institute of Science
and Technology
Ishikawa, Japan

Ag Asri Ag Ibrahim
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu, Sabah, Malaysia

Patricia Anthony
Lincoln University
Christchurch, New Zealand

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-13-2621-9 ISBN 978-981-13-2622-6 (eBook)
https://doi.org/10.1007/978-981-13-2622-6

Library of Congress Control Number: 2018955162

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Computational Science and Technology is a rapidly growing multi- and interdisciplinary field that uses advanced computing and data analysis to understand and solve complex problems. The absolute size of many challenges in computational science and technology demands the use of supercomputing, parallel processing, sophisticated algorithms and advanced system software and architecture. The ICCST 17 conference provides a unique forum to exchange innovative research ideas, recent results, and share experiences among researchers and practitioners in the field of advanced computational science and technology.

Building on the previous four conferences that include Regional Conference on Computational Science and Technology (RCSST 2007), the International Conference on Computational Science and Technology (ICCST 2014), the Third International Conference on Computational Science and Technology 2016 (ICCST 2016) and the Fourth International Conference on Computational Science and Technology 2017 (ICCST 2017) successful meetings, the FIFTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND TECHNOLOGY (ICCST 2018) pro-gram offers practitioners and researchers from academia and industry the possibility to share computational techniques and solutions in this area, to identify new issues, and to shape future directions for research, as well as to enable industrial users to apply leading-edge large-scale high-performance computational methods.

This volume presents a theory and practice of ongoing research in computational science and technology. The focuses of this volume is on a broad range of methodological approaches and empirical references points including artificial intelligence, cloud computing, communication and data networks, computational intelligence, data mining and data warehousing, evolutionary computing, high-performance computing, information retrieval, knowledge discovery, knowledge management, machine learning, modeling and simulations, parallel and distributed computing, problem-solving environments, semantic technology, soft computing, system-on-chip design and engineering, text mining, visualization and web-based and service computing . The carefully selected contributions to this volume were initially accepted for oral presentation during the Fifth International

Conference on Computational Science and Technology (ICCST18) held on 29–30th August in Kota Kinabalu, Malaysia. The level of contributions corresponds to that of advanced scientific works, although several of them could be addressed also to non-expert readers. The volume brings together 55 chapters.

In concluding, we would also like to express our deep gratitude and appreciation to all the program committee members, panel reviewers, organizing committees and volunteers for your fforts to make this conference a successful event. It is worth emphasizing that much theoretical and empirical work remains to be done. It is encouraging to find that more research on computational science and technology is still required. We sincerely hope the readers will find this book interesting, useful and informative and it will give then a valuable inspiration for original and innovative research.

Kota Kinabalu, Malaysia	Rayner Alfred
Ishikawa, Japan	Yuto Lim
Kota Kinabalu, Malaysia	Ag Asri Ag Ibrahim
Christchurch, New Zealand	Patricia Anthony

Contents

Automatic Classification and Retrieval of Brain Hemorrhages 1
Hau Lee Tong, Mohammad Faizal Ahmad Fauzi, Su Cheng Haw, Hu Ng and Timothy Tzen Vun Yap

Towards Stemming Error Reduction for Malay Texts 13
Mohamad Nizam Kassim, Shaiful Hisham Mat Jali, Mohd Aizaini Maarof and Anazida Zainal

Contactless Palm Vein ROI Extraction using Convex Hull Algorithm 25
Wee Lorn Jhinn, Michael Goh Kah Ong, Lau Siong Hoe and Tee Connie

A Robust Abnormal Behavior Detection Method Using Convolutional Neural Network 37
Nian Chi Tay, Tee Connie, Thian Song Ong, Kah Ong Michael Goh and Pin Shen Teh

Cryptanalysis of Improved and Provably Secure Three-Factor User Authentication Scheme for Wireless Sensor Networks 49
Jihyeon Ryu, Taeui Song, Jongho Moon, Hyoungshick Kim and Dongho Won

User Profiling in Anomaly Detection of Authorization Logs 59
Zahedeh Zamanian, Ali Feizollah, Nor Badrul Anuar, Miss Laiha Binti Mat Kiah, Karanam Srikanth and Sudhindra Kumar

Agent based integer programming framework for solving real-life curriculum-based university course timetabling 67
Mansour Hassani Abdalla, Joe Henry Obit, Rayner Alfred and Jetol Bolongkikit

3D Face Recognition using Kernel-based PCA Approach 77
Marcella Peter, Jacey-Lynn Minoi and Irwandi Hipni Mohamad Hipiny

Mobile-Augmented Reality Framework For Students Self-Centred Learning In Higher Education Institutions 87
Aaron Frederick Bulagang and Aslina Baharum

Computational Optimization Analysis of Feedforward plus Feedback Control Scheme for Boiler System 97
I. M. Chew, F. Wong, A. Bono, J. Nandong and K. I. Wong

Smart Verification Algorithm for IoT Applications using QR Tag 107
Abbas M. Al-Ghaili, Hairoladenan Kasim, Fiza Abdul Rahim,
Zul-Azri Ibrahim, Marini Othman and Zainuddin Hassan

Daily Activities Classification on Human Motion Primitives Detection Dataset 117
Zi Hau Chin, Hu Ng, Timothy Tzen Vun Yap, Hau Lee Tong,
Chiung Ching Ho and Vik Tor Goh

Implementation of Quarter-Sweep Approach in Poisson Image Blending Problem 127
Jeng Hong Eng, Azali Saudi and Jumat Sulaiman

Autonomous Road Potholes Detection on Video 137
Jia Juang Koh, Timothy Tzen Vun Yap, Hu Ng, Vik Tor Goh,
Hau Lee Tong, Chiung Ching Ho and Thiam Yong Kuek

Performance Comparison of Sequential and Cooperative Integer Programming Search Methodologies in Solving Curriculum-Based University Course Timetabling Problems (CB-UCT) 145
Mansour Hassani Abdalla, Joe Henry Obit, Rayner Alfred
and Jetol Bolongkikit

A Framework for Linear TV Recommendation by Leveraging Implicit Feedback 155
Abhishek Agarwal, Soumita Das, Joydeep Das and Subhashis Majumder

Study of Adaptive Model Predictive Control for Cyber-Physical Home Systems 165
Sian En Ooi, Yuan Fang, Yuto Lim and Yasuo Tan

Implementation of Constraint Programming and Simulated Annealing for Examination Timetabling Problem 175
Tan Li June, Joe H. Obit, Yu-Beng Leau and Jetol Bolongkikit

Time Task Scheduling for Simple and Proximate Time Model in Cyber-Physical Systems 185
Yuan Fang, Sian En Ooi, Yuto Lim and Yasuo Tan

DDoS Attack Monitoring using Smart Controller Placement in Software Defined Networking Architecture 195
Muhammad Reazul Haque, Saw C. Tan, Zulfadzli Yusoff, Ching K. Lee and Rizaludin Kaspin

Malay Language Speech Recognition for Preschool Children using Hidden Markov Model (HMM) System Training 205
Marlyn Maseri and Mazlina Mamat

A Formal Model of Multi-agent System for University Course Timetabling Problems 215
Kuan Yik Junn, Joe Henry Obit, Rayner Alfred and Jetol Bolongkikit

An Investigation towards Hostel Space Allocation Problem with Stochastic Algorithms 227
Joe Henry Obit, Kuan Yik Junn, Rayner Alfred, Jetol Bolongkikit and Ong Yan Sheng

Sensor Selection based on Minimum Redundancy Maximum Relevance for Activity Recognition in Smart Homes 237
Saed Sa'deh Juboor, Sook-Ling Chua and Lee Kien Foo

Improving Network Service Fault Prediction Performance with Multi-Instance Learning 249
Leonard Kok, Sook-Ling Chua, Chin-Kuan Ho, Lee Kien Foo and Mohd Rizal Bin Mohd Ramly

Identification of Road Surface Conditions using IoT Sensors and Machine Learning 259
Jin Ren Ng, Jan Shao Wong, Vik Tor Goh, Wen Jiun Yap, Timothy Tzen Vun Yap and Hu Ng

Real-Time Optimal Trajectory Correction (ROTC) for Autonomous Omnidirectional Robot 269
Noorfadzli Abdul Razak, Nor Hashim Mohd Arshad, Ramli Bin Adnan, Norashikin M. Thamrin and Ng Kok Mun

Incorporating Cellular Automaton based Microscopic Pedestrian Simulation and Genetic Algorithm for Spatial Layout Design Optimization 283
Najihah Ibrahim, Fadratul Hafinaz Hassan and Safial Aqbar Zakaria

QoE Enhancements for Video Traffic in Wireless Networks through Selective Packet Drops 295
Najwan Khambari and Bogdan Ghita

Preventing Denial of Service Attacks on Address Resolution in IPv6 Link-local Network: AR-match Security Technique 305
Ahmed K. Al-Ani, Mohammed Anbar, Selvakumar Manickam, Ayman Al-Ani and Yu-Beng Leau

Hybridizing Entropy Based Mechanism with Adaptive Threshold Algorithm to Detect RA Flooding Attack in IPv6 Networks 315
Syafiq Bin Ibrahim Shah, Mohammed Anbar, Ayman Al-Ani and Ahmed K. Al-Ani

Frequent Itemset Mining in High Dimensional Data: A Review 325
Fatimah Audah Md. Zaki and Nurul Fariza Zulkurnain

Validation of Bipartite Network Model of Dengue Hotspot Detection in Sarawak 335
Woon Chee Kok and Jane Labadin

Comparison of Classification Algorithms on ICMPv6-Based DDoS Attacks Detection 347
Omar E. Elejla, Bahari Belaton, Mohammed Anbar, Basim Alabsi and Ahmed K. Al-Ani

Feedforward plus Feedback Control Scheme and Computational Optimization Analysis for Integrating Process 359
I. M. Chew, F. Wong, A. Bono, J. Nandong and K. I. Wong

Performance Evaluation of Densely Deployed WLANs using Directional and Omni-Directional Antennas 369
Shuaib K. Memon, Kashif Nisar and Waseem Ahmad

Analyzing National Film Based on Social Media Tweets Input Using Topic Modelling and Data Mining Approach 379
Christine Diane Ramos, Merlin Teodosia Suarez and Edward Tighe

Perception and Skill Learning for Augmented and Virtual Reality Learning Environments 391
Ng Giap Weng and Angeline Lee Ling Sing

Application of Newton-4EGSOR Iteration for Solving Large Scale Unconstrained Optimization Problems with a Tridiagonal Hessian Matrix 401
Khadizah Ghazali, Jumat Sulaiman, Yosza Dasril and Darmesah Gabda

Detecting Depression in Videos using Uniformed Local Binary Pattern on Facial Features 413
Bryan G. Dadiz and Conrado R. Ruiz Jr.

Malicious Software Family Classification using Machine Learning Multi-class Classifiers 423
Cho Cho San, Mie Mie Su Thwin and Naing Linn Htun

Modification of AES Algorithm by Using Second Key and Modified SubBytes Operation for Text Encryption 435
Aye Aye Thinn and Mie Mie Su Thwin

Residential Neighbourhood Security using WiFi 445
Kain Hoe Tai, Vik Tor Goh, Timothy Tzen Vun Yap and Hu Ng

Prediction of Mobile Phone Dependence Using Bayesian Networks 453
Euihyun Jung

Learning the required entrepreneurial best practices using data mining algorithms 461
Waseem Ahmad, Shuaib K. Memon, Kashif Nisar and Gurpreet Singh

Agent Based Irrigation Management for Mixed-Cropping Farms 471
Kitti Chiewchan, Patricia Anthony and Sandhya Samarasinghe

A Review on Agent Communication Language 481
Gan Kim Soon, Chin Kim On, Patricia Anthony and Abdul Razak Hamdan

Simulation and Fabrication of Micro Magnetometer Using Flip-Chip Bonding Technique 493
Tengku Muhammad Afif bin Tengku Azmi and Nadzril bin Sulaiman

A Review on Recognition-Based Graphical Password Techniques 503
Amanul Islam, Lip Yee Por, Fazidah Othman and Chin Soon Ku

A Management Framework for Developing a Malware Eradication and Remediation System to Mitigate Cyberattacks 513
Nasim Aziz, Zahri Yunos and Rabiah Ahmad

A Review on Energy Efficient Path Planning Algorithms for Unmanned Air Vehicles 523
Sanjoy Kumar Debnath, Rosli Omar and Nor Badariyah Abdul Latip

Wireless Wearable for Sign Language Translator Device using Intel UP Squared (*UP*^{2}) Board 533
Tan Ching Phing, Radzi Ambar, Aslina Baharum, Hazwaj Mhd Poad and Mohd Helmy Abd Wahab

APTGuard: Advanced Persistent Threat (APT) Detections and Predictions using Android Smartphone 545
Bernard Lee Jin Chuan, Manmeet Mahinderjit Singh and Azizul Rahman Mohd Shariff

Smart Home using Microelectromechanical Systems (MEMS) Sensor and Ambient Intelligences (SAHOMASI) 557
Manmeet Mahinderjit Singh, Yuto Lim and Asrulnizam Manaf

Context Aware Knowledge Bases for Efficient Contextual Retrieval: Design and Methodologies 569
Sharyar Wani, Tengku Mohd, Tengku Sembok
and Mohammad Shuaib Mir

Correction to: A Review on Energy Efficient Path Planning Algorithms for Unmanned Air Vehicles C1
Sanjoy Kumar Debnath, Rosli Omar and Nor Badariyah Abdul Latip

Author Index 581

Automatic Classification and Retrieval of Brain Hemorrhages

Hau Lee Tong[1], Mohammad Faizal Ahmad Fauzi[2], Su Cheng Haw[1], Hu Ng[1] and Timothy Tzen Vun Yap[1]

[1]Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
[2]Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
hltong@mmu.edu.my

Abstract. In this work, Computed Tomography (CT) brain images are adopted for the annotation of different types of hemorrhages. The ultimate objective is to devise the semantics-based retrieval system for retrieving the images based on the different keywords. The adopted keywords are hemorrhagic slices, intra-axial, subdural and extradural slices. The proposed approach is consisted of three separated annotation processes are proposed which are annotation of hemorrhagic slices, annotation of intra-axial and annotation of subdural and extradural. The dataset with 519 CT images is obtained from two collaborating hospitals. For the classification, support vector machine (SVM) with radial basis function (RBF) kernel is considered. On overall, the classification results from each experiment achieved precision and recall of more than 79%. After the classification, the images will be annotated with the classified keywords together with the obtained decision values. During the retrieval, the relevant images will be retrieved and ranked correspondingly according to the decision values.

Keywords: Brain Hemorrhages, Image Classification, Image Retrieval.

1 Introduction

Intracranial hemorrhage detection is clinically significant for the patients having head trauma and neurological disturbances. This is because early discovery and accurate diagnosis of the brain abnormalities is crucial for the execution of the successful therapy and proper treatment. Multi-slice CT scans are extensively utilized in today's analysis of head traumas due to its effectiveness to unveil some abnormalities such as calcification, hemorrhage and bone fractures.

A lot of research works have been carried out to assist the visual interpretation of the medical brain images. Brain hemorrhage can cause brain shift and make it lose its symmetric property. As such, investigation of the symmetric information can assist in hemorrhage detecting. Chawla et al. proposed the detection of line of symmetry based on the physical structure of the skull [1]. Likewise, Saito et al. proposed a method to detect the brain lesion based on the difference values of the extracted features be-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_1

tween region of interests (ROIs) at symmetrical position [2]. The symmetric line is determined from the midpoints of the skull contour. Besides the symmetric detection approach, segmentation technique is commonly adopted to obtain the potential desired regions. Bhadauria and Dewal proposed an approach by combining the features of both fuzzy c-means and region-based active contour method to segment the hemorrhagic regions [3]. Al-Ayyoub et al. used Otsu's thresholding method for the image segmentation [4]. From the Otsu's method, the potential ROIs are obtained. Jyoti et al. proposed genetic FCM (GFCM), a clustering algorithm integrated with Genetic Algorithm (GA) to segment the CT brain images with the modified objective function [5]. Suryawanshi and Jadhao adopted a neural network approach with watershed segmentation to detect the intracerebral haemorrhage and subdural [6].

Huge collections of medical images have contributed significant source of knowledge and various areas of medical research. However, the major problem is that the size of the medical image collection in hospitals faces constant daily growth. Thus the amount of work and time required for retrieving a particular image for further analysis is costly. As such, image retrieval particularly in medical domain has gained some attention in the research area of content-based image retrieval (CBIR) [7-10].

Consequently, multi-slice CT brain images are adopted in this work to classify, annotate and retrieve the images by using different keywords based on the type of hemorrhages. The retrieval is based on the adopted keywords which are hemorrhagic slices, intra-axial, subdural and extradural slices

2 Proposed System Overview

The architecture diagram of our proposed system is shown in Fig. 1. The proposed system consists of two main components, i.e., the offline mode and online mode. The offline component will be the automated annotation part. Each CT scan will be annotated by the keywords and the attached semantic keyword will be stored into a database. On the other hand, the online component will be the retrieval part. CT images can be retrieved according to the predefined keywords.

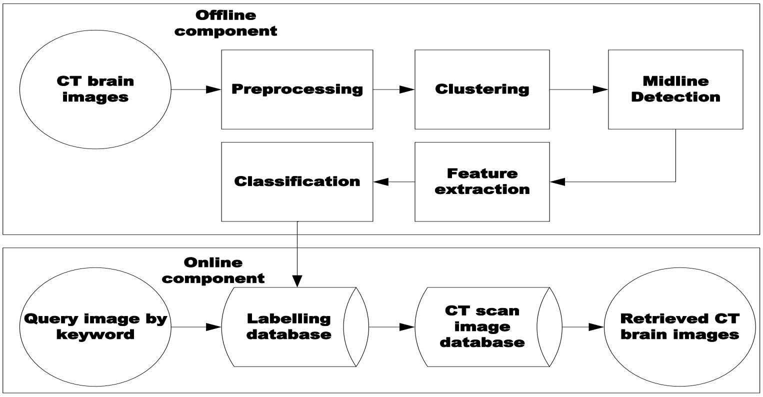


Fig. 1. Overview of proposed annotation and retrieval

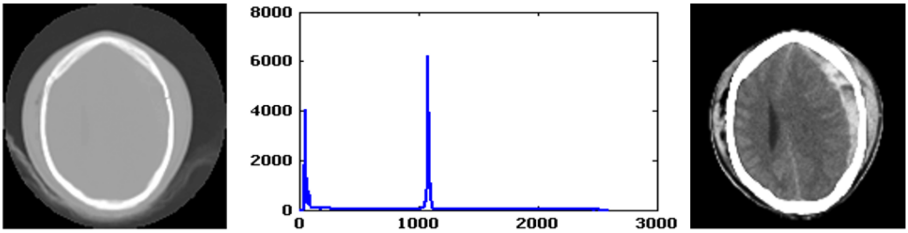
3 Preprocessing

3.1 Original Image Enhancement

The original CT images are lacking in dynamic range whereby only limited objects are visible. The major aim of the enhancement is to augment the interpretability and perception of images to contribute better input for the subsequent image processing process. Firstly the histogram for the original image is constructed as shown in Fig. 2(a). The constructed histogram is consisted of several peaks. However, only the rightmost peak is within the significant range for region of interest which is the intracranial area. Then, the curve smoothing process is conducted where convolution operation with a vector is applied. Each element of the vector is of the value 10^{-3}. The smoothened curve is converted into absolute first difference (ABS) in order to generate two highest peaks. The two generated peaks will be the lower limit, I_L and upper limit, I_U. The acquired I_L and I_U are utilized for the linear contrast stretching as shown in equation (1).

$$F(i,j) = I_{\max} \frac{(I(i,j) - I_L)}{(I_U - I_L)} \tag{1}$$

where Imax, $I(i,j)$ and $F(i,j)$ denote the maximum intensity of the image, pixel value of the original image and pixel value of the contrast improved image respectively. After contrast stretching, the contrast enhanced image is shown in Fig. 2(b).



(a) (b) (c)

Fig. 2. (a) Original Image (b) Histogram of the original image (b) Enhanced Image

3.2 Parenchyma Extraction

The subsequent preprocessing is to extract the parenchyma from the enhanced image. In order to obtain the parenchyma, thresholding technique is employed to isolate the background, skull and scalp from it. Usually, the skull always appears to be the largest connected component compared with the objects in the background. Therefore, the largest connected component is identified in order to detect the skull. Subsequently, parenchyma mask is generated by filling up the hole inside the skull. Finally, intensity of the skull is set to zero and the acquired parenchyma is shown in Fig. 3.

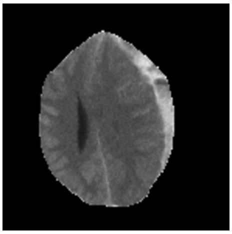


Fig. 3. The acquired parenchyma

3.3 Potential Hemorrhagic Regions Contrast Enhancement

The purpose of third preprocessing is to further enhance the parenchyma for the subsequent clustering stage. Thus, this enhancement aim to make the hemorrhagic regions more visible to clearly reveal the dissimilarity between the hemorrhagic hemisphere and non-hemorrhagic hemisphere. Prior to contrast stretching, the appropriate lower and upper limits need to be obtained. Firstly construct the histogram for the acquired parenchyma. Then identify the lower limit, I_L which is the peak position of the constructed histogram. From the obtained lower limit, the upper limit can be derived by equation (2). The determination of the appropriate lower and upper limits is necessary to ensure the focus of the stretching is for the hemorrhagic regions rather than normal regions.

$$I_U = I_L + Ie \tag{2}$$

where *I*e is predefined at 500 as found from experimental observation.

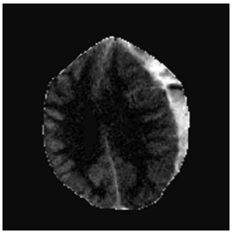


Fig. 4. Hemorrhagic region contrast enhanced image

Lastly, input the auto-determined values of I_L and I_U into equation (1) for the contrast stretching and the result is depicted in Fig. 4.

4 Potential Hemorrhagic Region Clustering

The main goal for this section is to garner potential hemorrhagic regions into a single cluster to be used for the annotation of the hemorrhages. Hemorrhagic regions always appear as bright regions. Therefore only intensity of the images is considered for the clustering. In order to achieve this, firstly, the image is partitioned into two clusters. From these two clusters, the low intensity cluster without potential hemorrhagic regions is ignored. Only the high intensity cluster which consists of potential hemorrhagic regions is considered. Four clustering techniques which are Otsu thresholding, fuzzy c-means (FCM), k-means and expectation-maximization (EM) are attempted in order to select the most appropriate technique for subdural, extradural and intra-axial hemorrhages annotation.

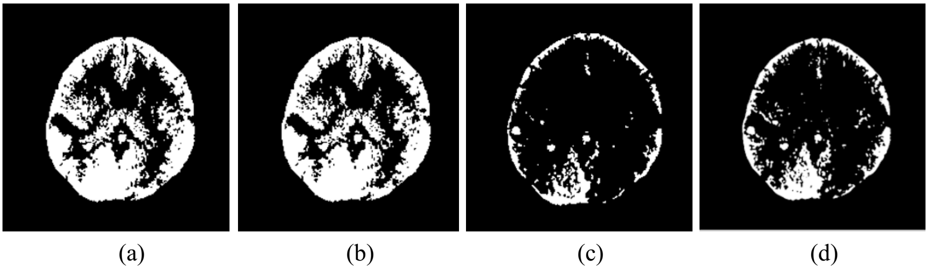


Fig. 5. Clustering results by (a) Otsu thresholding (b) FCM clustering (c) K-means clustering and (d) EM clustering

From the results obtained as shown in Fig. 5, Otsu thresholding, FCM and EM encountered over-segmentation as hemorrhagic region is merged together with neighbouring pixels and more noises are present. On the other hand, k-means conserves the appropriate shape most of the time and yields less noise. This directly contributes to more precise shape properties acquisition at the later region-based feature extraction. Consequently, k-means clustering is adopted in our system.

5 Midline detection

The proposed approach of the parenchyma midline acquisition consists of two stages. Firstly, acquire the contour of the parenchyma as shown in Fig. 6(a). Then the line scanning process will be executed: The scanning begins from the endpoints of the top and bottom sub-contours to locate the local minima and local maxima for top and bottom sub-contour respectively. The line scanning of local maxima, (x_1,y_1) for bottom sub-contour is as illustrated in Fig. 6(d).

In the case where the local minima or maxima point is not detected, Radon transform [11] will be utilized to shorten the searching line. Radon transform is defined as:

$$R(\rho,\theta) = \iint f(x,y)\delta(\rho - x\cos\theta - y\sin\theta)dxdy \tag{3}$$

where ρ,θ and $f(x,y)$ denote distance from origin to the line, angle from the X-axis to the normal direction of the line and pixel intensity at coordinate (x,y). Dirac delta function is the obtained shortened line is thickened as depicted in Fig. 6(e).

Then moving average is applied to locate the points of interest, (x_2,y_2) from the shortened contour line. Basically moving average is used for the computation of the average intensity for all the points along the contour line. The location of the points is based on the highest average value of intensity as shown in Fig. 6 (f).

Eventually, midline is formed by using the linear interpolation as defined in equation (4). The detected midlines are shown in Fig. 7.

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)} \tag{4}$$

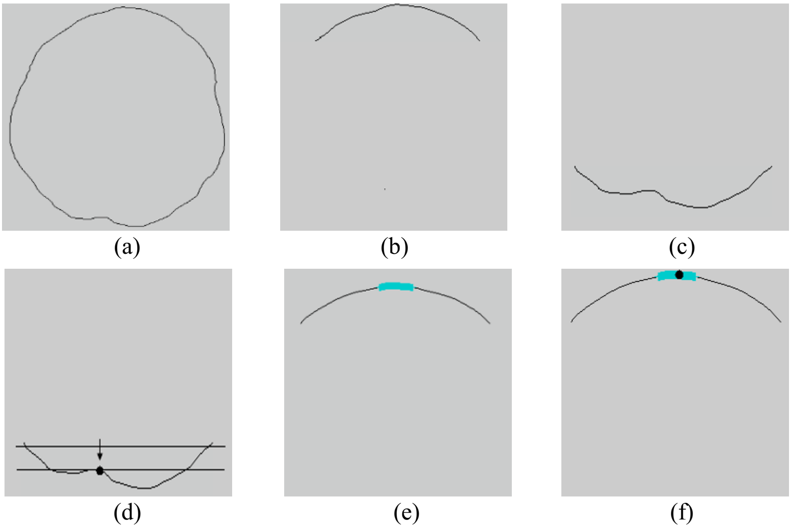


Fig. 6. (a) Contour of parenchyma area (b) Top sub-contour (c) Bottom sub-contour (d) Line scanning for local maxima detection (e) Shortened searching line (f) Located highest average value of intensity point

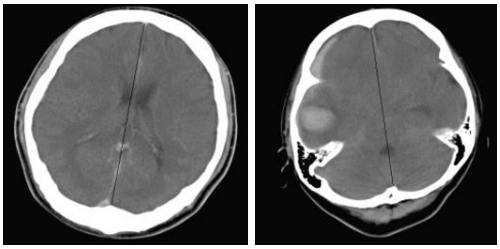


Fig. 7. The detected midline

6 Feature Extraction

Features are extracted from the left and right hemispheres and used to classify the images. There is a three-stage feature extraction which is feature extraction for hemorrhagic slices, feature extraction for intra-axial slices and feature extraction for subdural and extradural. Twenty three features are proposed based on their ability to distinguish hemorrhagic and non-hemorrhagic slices. These twenty three features are derived based on one feature from Local Binary Pattern (LBP), one from entropy, four from four-bin histogram, one from intensity histogram and sixteen from four Haralick texture features(energy, entropy, autocorrelation and maximum probability) descriptors on four directions. Prior to the feature extraction for intra-axial, subdural and extradural, the potential hemorrhagic regions will be divided into intra regions and boundary regions. Intra regions will proceed with the intra-axial feature extraction with the twenty three that identical with the features of hemorrhagic slices. On the other hand, the boundary regions will proceed with the region-based feature extraction to annotate the subdural and extradural. The twelve shape features adopted are region area, border contact area, orientation, linearity, concavity, ellipticity, circularity, triangularity, solidity, extent, eccentricity and sum of centroid contour distance curve Fourier descriptor(CCDCFD). Some of the adopted shape features are defined in equation (4) to equation (8)

Linearity, $$L(ROI) = 1 - \frac{\text{minor axis}}{\text{major axis}} \tag{5}$$

Triangularity, $$\angle(ROI) = \begin{cases} 108I & if\ I \le \dfrac{1}{108}, \\ \dfrac{1}{108I} & otherwise \end{cases} \tag{6}$$

Where $I = \dfrac{\mu_{2,0}(ROI)\mu_{0,2}(ROI) - (\mu_{1,1}(ROI))^2}{(\mu_{0,0}(ROI))^4}$ and $\mu_{i,j}(ROI)$ is the (i,j)-moment

Ellipticity, $$1(ROI) = \begin{cases} 16\pi^2 I & if\ I \le \dfrac{1}{16\pi^2}, \\ \dfrac{1}{16\pi^2 I} & otherwise \end{cases} \tag{7}$$

Circularity, $$\partial(ROI) = \frac{(\mu_{0,0}(ROI))^2}{2\pi(\mu_{2,0}(ROI) + \mu_{0,2}(ROI))} \tag{8}$$

Sum of CCDC= $$\sum_{k=2}^{N-1} \left| \frac{ft(k)}{ft(1)} \right| \tag{9}$$

Where $ft(k) = \dfrac{1}{N}\sum_{n=0}^{N-1} z(n)\exp\left(\dfrac{-i2\pi kn}{N}\right)$, k=0, 1, 2…., N-1, z(n)=$x_1(n)$ $+iy_1(n)$ and

contour points= (x_1(n) , y_1(n)).

Region area, border contact area and orientation are employed to differentiate the normal regions from the subdural and extradural regions. The subsequent eight features are primarily adopted to differentiate extradural from subdural. Extradural and subdural always appear to be bi-convex and elongated crescent in shape respectively. Generally, extradural is more solid, extent, elliptic, circular and triangular as compared to subdural. However, subdural is more concave and linear than extradural.

7 Classification

The acquired CT brain images are in DICOM format with their dimension 512 x 512. The images are acquired from two collaborating hospitals, which are Serdang Hospital and Putrajaya Hospital, to test the feasibility of the proposed approach on datasets generated from different scanners and different setup. The dataset consists of 181 normal slices, 209 intra-axial slices, 60 extradural slices and 86 subdural slices. The adopted classification technique is SVM with RBF kernel. During the classification, ten-fold cross validation method is performed. The features obtained from three separate feature extraction process are channel into the classifier respectively in order to categorize the different slices. The obtained recall and precision are shown in Table 1. The recall and precision for all slices achieved satisfactory results with at least 0.79 for all. This is contributed by the features employed as they well describe the characteristics of the different slices. However, the recall for intra-axial slice is relatively lower compared with other slices. Some bright normal regions presented in non-intra-axial slices increase the similarity of the features between some non-intra-axial and intra-axial slices. Therefore, it generated the higher misclassification of intra-axial slices.

Table 1. Recall and precision for different types of slices.

	Recall	Precision
Normal slice	0.917	0.834
Hemorrhagic slice	0.902	0.953
Intra axial slice	0.793	0.925
Extradural	0.850	0.927
Subdural	0.892	0.858

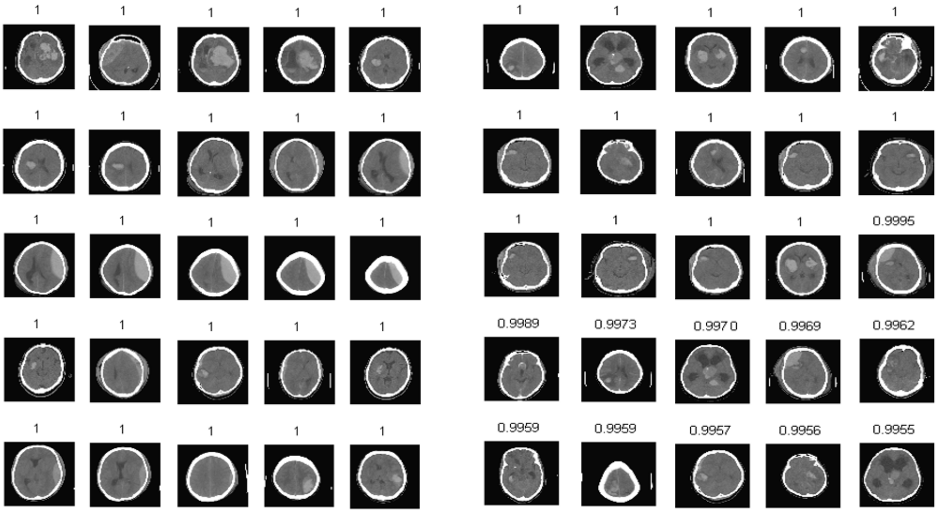
8 Retrieval Results

This section presents the retrieval results based on the keywords which are “hemorrhage”, “intra-axial”, “subdural” and “extradural”. The retrieval results for each keyword are exhibited based on the twenty five most relevant as shown in Fig. 8. The relevancy or ranking is based decision values obtained from RBF SVM.

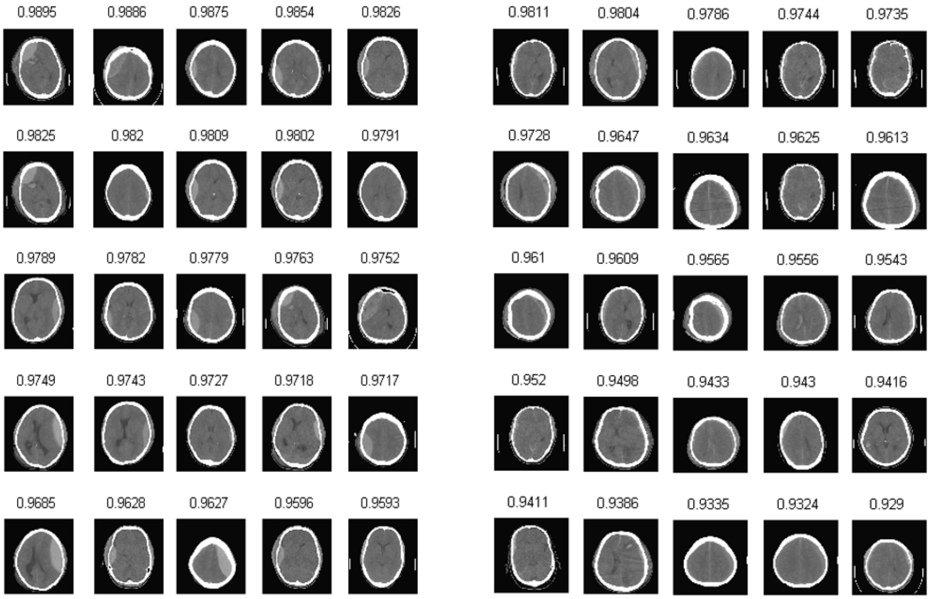
Besides the visual retrieval results, the accuracy of the retrieval is evaluated based on the numbers of most relevant to the adopted keywords over 519 images. The testing numbers for each keyword are different as the evaluation is according to the number of total slices of each keyword. From Table 2, the retrieval result based on the keyword “hemorrhage” is 96% for the 300 most relevant retrievals. In this case, there are 12 irrelevant images being retrieved together in the top 300. On the other hand, for the 100 most intra-axial relevant results, there are three irrelevant images being retrieved. For 50 most extradural relevant retrieval results, the accuracy obtained is 96%. There are 2 irrelevant images presented in the retrieval results. Lastly for the 50 most subdural relevant retrieval results, the accuracy obtained is 92%. There are 4 irrelevant images being retrieved in this retrieval.

9 Conclusion

This research work proposed three segregated feature extraction processes to classify and retrieve the different types of slices based on their different features and locations. By individualizing each of the process, the appropriate approach such as midline approach, global-based and region-based feature extractions can be adopted at each level for the categorization of specific hemorrhage. Overall, the precision and recall rates are over 79% for all the classification results. In our future work, we would like to extend classification types and retrieval keywords to include more abnormalities of brain such as infarct, hydrocephalus and atrophy.



(a) Hemorrhage (b) Intra-axial



(c) Extradural (d) Subdural

Fig. 8. Twenty five most relevant retrieval results

Table 2. Retrieval accuracy by using different keywords

Numbers of most relevant retrieval	Hemorrhage	Intra-axial	Extradural	Subdural
25	100.0%	100.0%	100.0%	100.0%
50	98.0%	96.0%	96.0%	92.0%
100	99.0%	97.0%		
200	99.5%			
300	96.0%			

References

1. Chawla, M., Sharma, S., Sivaswamy, J., Kishore, L. T.: A method for automatic detection and classification of stroke from brain ct images. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3581-3584. (2009).
2. Saito, H., Katsuragawa, S., Hirai, T., Kakeda, S., Kourogi, Y.: A computerized method for detection of acute cerebral infarction on ct images. Nippon Hoshasen Gijutsu Gakkai Zasshi 66(9), 1169–1177.(2010).
3. Bhadauria, H. S., Dewal, M. L.: Intracranial hemorrhage detection using spatial fuzzy c-mean and region-based active contour on brain CT imaging. Signal, Image and Video Processing 8(2), 357-364 (2014).

4. Al-Ayyoub, M., Alawad, D., Al-Darabsah, K., Aljarrah, I.: Automatic detection and classification of brain hemorrhages. WSEAS Transactions on Computers 12(10), 395-405 (2013).
5. Jyoti, A., Mohanty, M., Kar, S., Biswal, B.: Optimized clustering method for CT brain image segmentation. Advances in Intelligent Systems and Computing 327(1),317-324. (2015).
6. Suryawanshi, S. H.,Jadhao, K. T.: Smart brain hemorrhage diagnosis using artificial neural networks. International Journal of Scientific and Technology Research 4(10), p267-271 (2015).
7. Müller, H., Deserno, T.: Content-based medical image retrieval. Biomedical image processing, biological and medical physics, biomedical engineering, pp. 471–494. (2011).
8. Ramamurthy, B., Chandran, K., Aishwarya, S., Janaranjani, P.: CBMIR: content based image retrieval using invariant moments, glcm and grayscale resolution for medical images. European Journal of Scientific Research 59(4), 460-471 (2011).
9. Müller, H., de Herrera, A., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. CLEF Online Working Notes, pp. 1–16. (2012).
10. de Herrera, A. G., Kalpathy-Cramer, J., Fushman, D. D., Antani, S., Muller, H.: Overview of the ImageCLEF 2013 medical tasks. Working notes of CLEF, pp. 1-15. (2013).
11. Chen, B., Zhong, H.: line detection in image based on edge enhancement. Second International Symposium on Information Science and Engineering, pp. 415-418. IEEE Computer Society Washington, DC, USA (2009).

Towards Stemming Error Reduction for Malay Texts

Mohamad Nizam Kassim[1], Shaiful Hisham Mat Jali[2], Mohd Aizaini Maarof[2] and Anazida Zainal[2]

[1] CyberSecurity Malaysia, 43300 Seri Kembangan, Selangor
[2] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai Johore, Malaysia
nizam@cybersecurity.my shaiful.hisham@outlook.com
{aizaini,anazida}@utm.my

Abstract. Text stemmer is one of useful language preprocessing tools in the field of information retrieval, text mining and natural language processing. It is used to map morphological variants of words into base forms. Most of the current text stemmers for the Malay language focused on removing affixes, clitics, and particles from affixation words. However, these stemmers still suffered from stemming errors due to insufficiently address the root cause of these stemming errors. This paper investigates the root cause of stemming errors and proposes stemming technique to address possible stemming errors. The proposed text stemmer uses affixes removal method and multiple dictionary lookup to address various root causes of stemming errors. The experimental results showed promising stemming accuracy in reducing various possible stemming errors.

Keywords: Text Stemmer, Text Stemming, Stemming Errors, Stemming Algorithm, Rule-based Affixes Elimination

1 Introduction

Text stemming is a linguistic process in which various morphological variants of words (derived words) are mapped to their base forms (sometimes, called root words or stemmed words) [1]. It has been commonly used as language preprocessing tool in various applications such as information retrieval, text classification, text clustering, natural language and machine translation [2,3]. A text stemmer is developed based on microlinguistics such as morphology, syntax, and semantics. Every natural language has many word variants by adding morphological morpheme of affixes, clitics or particles to their base forms [4]. It is a great challenge to design and develop an effective text stemmer for morphologically rich languages is due to large numbers of morphological variants of word patterns. For instance, the English words *connects, connected, connection, connecting,* and *connector* are mapped to their base form *connect* with the help of text stemmer. On other hand, Malay words of *disambung* (connected), *menyambung* (connects), *sambungan* (connection), *penyambungan* (connection or continuation) and *kesinambungan* (continuity) are mapped to their base form *sambung* which different from English language due to four different structures of affixa-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_2

tion word patterns and various bound morphemes of affixes, clitics and particles attached to those word patterns. Hence, many current text stemmers were developed based on various stemming approaches such as the selection of longest/shortest affixes, clitics or particles, ordering affixation stemming rules, ordering lookup dictionaries and dictionary entries [5,6]. Despite the fact that the wide variety of text stemming approaches were developed, the current text stemmers for the Malay language still suffered from stemming errors. Therefore, there is a need for an enhanced stemming approach that improves stemming accuracy. Hence, this paper proposes an enhanced text stemmer that combines affixes removal method and dictionary lookup to address possible stemming errors. This paper is organized into five sections. Section 2 highlights the related works of the current text stemmers. Section 3 describes the linguistic aspects of the Malay language. Section 4, describes an enhanced text stemming approach to address possible stemming errors. Section 5 discusses experimental results and explains our findings with respect to the proposed text stemmer. Finally, Section 6 concludes this paper with a brief summary.

2 Related Works

The early text stemmer was developed by Lovins [8] for the English language and inspired other researchers to develop text stemmers in various natural languages such as Latin, Dutch, and Arabic. This early text stemmer has also led to active research on the development of other improved text stemmers for the English language such as Porter's stemmer, Paice/Husk’s stemmer and Hull’s stemmer [3]. The Porter’s stemmer is considered as the most prominent text stemmer due to its application in information retrieval. It has become a defacto text stemmer for the English language and the same stemming approaches used in Porter’s stemmer have been adopted for use in other natural languages such as the Romance (French, Italian, Portuguese and Spanish), Germanic (Dutch and German), Scandinavian (Danish, Norwegian and Swedish), Finnish and Russian languages [3]. Similarly, the early text stemmer for the Malay language was developed Othman [9] and also influenced other researchers to develop an improved text stemmer for the Malay language. The subsequent text stemmers applied various stemming approaches to stem affixation words. It is important to highlight that past researchers faced various stemming errors in developing a perfect text stemmer for the Malay language [6]. This scenario led to various stemming approaches has been developed to address these stemming errors. Currently, six different text stemming approaches were proposed in the current text stemmers for Malay language, i.e., rule-based text stemming [9,10], Rules Application Order (RAO) [5,11,12], Rules Frequency Order (RFO) [13], Modified RFO [7,14], Modified Porter Stemmer [15], and Syllable-Based text stemming [16]. There are two different strategies for text stemming research, i.e., first, rule-based text stemming and second, other text stemming. The first text stemming strategies were to improve upon the rule-based text stemmer developed by Othman [9], Rule-based stemming applies sets of affixation text stemming rules in alphabetical order and dictionary lookup for checking that a word is not the root word prior to applying text stemming. However, Sembok et al. [17] highlighted the weaknesses of the rule-based stemming approach in which produced many stemming errors. Consequently, the RAO stemming approach was introduced, which adds and rearranges affixation text stemming rules in morphological order and also uses a dictionary lookup to check that the input word is not the root word before and after applying text stemming [7]. Based on the RAO stemming approach, the RFO stemming approach was then introduced, which considered only the most frequent affixes in Malay texts and used a dictionary lookup to check that the word is not the root word before and after applying text stemming rules [13]. Finally, Modified RFO stemming was introduced, which uses two types of dictionaries: root word dictionaries and 'background knowledge' derivative word dictionaries [14]. The second text stemming strategies were adopting other stemming approaches such as the Modified Porter Stemmer [15] and Syllable-Based methods [16]. However, these stemming approaches were not popular with researchers due to the complexity of Malay morphology, which contains spelling variations and exceptions rules wherein affix removal is not sufficient for finding root words and recoding the first letter is required to produce the correct root words. In short, these text stemming approaches were developed to improve stemming accuracy to stem affixation words in Malay texts.

3 Malay Language

In general, the Malay language is an Austronesian language that is widely spoken in South East Asia region and has very complex morphological structures [4]. Unlike other languages such as the English language, the Malay language has a large number of morphological variants of the words. The understanding of morphological structure of Malay language is an important consideration for the development of a text stemmer because the text stemmer maps various morphological variants of the word to their base forms. Therefore, it is important to understand the morphological processes by which base forms evolve to derived words. In the Malay language, derived words may contain a combination of possible morphemes (called affixes) that attached to root word such as prefixes (e.g. *di+, per+, peng+*)*,* suffixes (e.g. *+i, +kan, +an*), confixes (e.g. *dike+i, kepel+an, meng+kan)* and infixes (e.g. *+el+, +er+, +em+*). These possible combinations made up affixation words in the Malay language, i.e., prefixation, suffixation, confixation and infixation words [2]. Other possible combinations are root words are attached with proclitics (e.g. *ku+, kau+*), enclitics (e.g. *+mu, +nya)*, or particles (e.g. *+lah, +kah).* There are various possible combinations of root words with prefixes, suffixes, confixes, infixes, proclitics, enclitics and particles that lead to the formation of prefixation (e.g. *bermain, mempersenda*), suffixation (e.g. *kenalan, kenalanmu*), confixation (e.g. *permainan, kesudahannya*, *mempertemukannya*, *kunantikannya*), and infixation (e.g. *telunjuk*) in the Malay language. There are also special variation and exception rules in deriving prefixation and confixation words whereby prefixes e.g. *meny+, pem+* and *pen+* require first letter of root word to be dropped e.g. *meny + samun* (steal) → *menyamun (stealing)*, *pem + pilih* (choose) → *pemilih* (choosing) and *pen + tolong* (assist) → *penolong* (assistant).

4 Understanding Root Causes of Stemming Errors

Various possible affixation stemming errors have been associated with the current Malay text stemmers [5,7,11-12, 16]. Researchers categorize these stemming errors in terms of overstemming, understemming, unstem and spelling variations and exceptions errors. The root causes of these stemming errors will be further discussed in this section.

4.1 Overstemming Error Due To Root Words With Similar Morphemes To Affixes in Affixation Words

Root words with similar morphemes to affixes in affixation words cause stemming errors in the current Malay text stemmers. In general, these root words are incorrectly stemmed as affixation words, which leads to overstemming errors. The following examples illustrate these overstemming errors:

Scenario I: Root words with similar morphemes to prefixation, suffixation and confixation words

- *berat* (heavy) → *rat* when the prefix (*be+*) is removed (overstemming)
- *pasukan* (team) → *pasu* when the confix (*ber+an*) is removed (overstemming)
- *menteri* (minister) → *ter* when the confix (*ber+an*) is removed (overstemming)

To address these stemming errors, the current Malay text stemmers adopt a dictionary lookup to accept these root words as root words. Thus, these root words are not stemmed by affixation text stemming rules. However, it is challenging to apply text stemming rules to affixation words that contain these root words, as the following examples illustrate:

Scenario I: Affixation words containing root words with similar morphemes to prefixation, suffixation and confixation words

- *berkesan* (effective) → *kes* when the confix (*ber+an*) is removed (overstemming)
- *pasukannya* (his/her team) → *pasu* when the suffix (*+an*) and enclitics (*+nya*) are removed (overstemming)
- *menterinya* (minister) → *ter* when the confix (*men+i*) and enclitics (*+nya*) are removed (overstemming)

Unfortunately, the use of a root word dictionary lookup only matches each word in text documents if the word is a root word as opposed to an affixation word. Hence, the proposed text stemmer will address these stemming errors by using the root word and derivative dictionaries. The root word dictionaries accept root words with similar morphemes to affixes in affixation words as root words whereas the derivative dictionaries stem affixation words that contain root words with similar morphemes to affixes in affixation words (e.g. *berkesan, pasukannya, menterinya*) to find the correct root words.

4.2 Overstemming or Understemming Error Due To Affixation Word With Two Possible Root Words

Root word ambiguity has not been addressed in current Malay text stemmers, as highlighted by Darwis *et al.* [7]. The root word ambiguity problem occurs when there is more than one root word that can be selected during text stemming. The main problem associated with root word ambiguity involves selecting the correct root word from affixation word that contains multiple root words. For example, affixation words may contain two possible root words, depending on the type of affix matching selected, which could lead to stemming errors as follows:

Scenario I:

- Incorrect: *beribu* (thousands) → *ibu* (mother) when the prefix (*ber*+) is removed
- Correct: *beribu* (thousands) → *ribu* (thousand) when the prefix (*be*+) is removed

Scenario II:

- Incorrect: *katakan* (let say) → *katak* (frog) when the suffix (+*an*) is removed
- Correct: *katakan* (let say) → *kata* (word) when the suffix (+*kan*) is removed

Scenario III:

- Incorrect: *perangkaan* (statistics) → *rangka* (skeleton) when the confix (*pe*+*an*) is removed
- Correct: *perangkaan* (statistics) → *angka* (number) when the confix (*per*+*kan*) is removed

Unfortunately, the current Malay text stemmers use of a root word dictionary to check against stemmed words being valid root words in which do not address these stemming errors. Thus, the proposed text stemmer will use derivative dictionaries to stem these words to avoid root word ambiguity.

4.3 Overstemming or Understemming Errors Due To Affixation Word That Has Two Possible Affixes Matching Selection

Selecting the type of affix matching to use to identify the affixes of affixation words that need to be removed is very crucial in text stemming. There are two types of affix matching, i.e., longest affix matching and shortest affix matching. Longest affix matching fits the longest to shortest affixes of affixation words whereas shortest affix matching fits the shortest to longest affixes of affixation words. However, current research suggests that both types of affix matching lead to affixation stemming errors for specific affixation words [5], as shown in the following example:

Scenario I: Affix matching selection in prefixation text stemming

- Incorrect: *berasa* (to feel) → *asa* when longest affix matching is selected
- Correct: *berasa* (to feel) → *rasa* when shortest affix matching is selected
- Incorrect: *beranak* (to give birth) → *ranak* when shortest affix matching is selected
- Correct: *beranak* (to give birth) → *anak* when longest affix matching is selected

Scenario II: Affix matching selection in suffixation text stemming

- Incorrect: *ajakannya* (invitation) → *aja* when longest affix matching is selected
- Correct: *ajakannya* (invitation) → *ajak* when shortest affix matching is selected
- Incorrect: *biarkanlah* (let it be) → *biark* when shortest affix matching is selected
- Correct: *biarkanlah* (let it be) → *biar* when longest affix matching is selected

To address these stemming errors, the proposed text stemmer will select the longest affix match in its affixation text stemming rules and use derivative dictionaries to reduce affixation stemming errors.

4.4 Overstemming or Understemming Errors Due To Spelling Variation And Exception in Prefixation And Confixation Words

Unlike other natural languages such as English and French, the Malay language has very rich morphological rules involving two steps of affix removal: first, the affixes are removed and then second, the affixes are replaced with single letters. The text stemming challenges arise when very similar affixation words, i.e., *memilih* (to choose), *memikir* (to think) and *meminum* (to drink), are stemmed. The potential for special variation and exception stemming errors for highly similar affixation words is illustrated by the following examples:

Scenario I

- Correct: *memilih* → *pilih* by removing prefix (*mem*+) and inserting character *p*
- Incorrect: *memilih* → *filih* by removing prefix (*mem*+) and inserting character *f*
- Incorrect: *memilih* → *milih* by removing prefix (*me*+)

Scenario II

- Correct: *memikir* → *fikir* by removing prefix (*mem*+) and inserting character *f*
- Incorrect: *memikir* → *pikir* by removing prefix (*mem*+) and inserting character *p*
- Incorrect: *memikir* → *mikir* by removing the prefix (*me*+)

Scenario III

- Correct: *meminum* → *minum* by removing prefix (*me*+)
- Incorrect: *meminum* → *finum* by removing prefix (*mem*+) and inserting character *f*
- Incorrect: *meminum* → *pinum* by removing prefix (*mem*+) and inserting character *p*

To address these stemming errors, the proposed text stemmer differentiates conflicting morphological rules for spelling variations and exceptions by using affixation stemming rules and multiple derivative dictionaries. Therefore, the proposed text stemmer would not have difficulties selecting the correct affixes to remove from affixation words, whereas the current text stemmers only depend on affixation stemming rules and a root word dictionary.

In short, the proposed text stemmer has six main modules: Input module, Word Checking module, Word Processing module, Dictionary-based Text stemming module, Rule-based Text stemming module and the Output module, as described in Algorithm 1.

Algorithm 1. The proposed text stemmer

```
Input: Accept the input text document and remove special characters
       Convert all words from upper case to lower case and go to Step-1

Output: Root Words {stem1, stem2, stem3...stemn}

Step-1 Word Checking: i = word1, word2, word3...wordn
IF i = 0, go to Output
IF i = wordn, go to Step-2

Step-2 Checking Dictionary Lookup for Non-Derived Words
Check wordn against non-derived words (root words)
IF wordn = root words and proper nouns, accept wordn as root word and go to Step-1
ELSE go to Step-3

Step-3 Dictionary-based Text Stemming for Affixation Words
Check wordn against affixation words for rare word patterns
IF wordn = infixation, stem wordn and go to Step-1
ELSE go to Step-4
Check wordn against affixation words for conflicting morphological rules
IF wordn = confixation, prefixation and suffixation, stem wordn and go to Step-1
ELSE go to Step-4

Step-4 Rule-based Text Stemming for Affixation Words
Check wordn against affixation words for simple morphological rules
IF wordn = confixation, prefixation and suffixation, stem wordn and go to Step-1
ELSE accept wordn as root word and go to Step-1
```

Each module has specific functions in the proposed text stemmer to address possible stemming errors in Malay texts. The first module, the Input module, accepts a text document under specific character encoding (plain text) and pre-processes it to remove unnecessary special characters before any text stemming processes are performed by the subsequent modules. The second module, Word Checking module, checks each word from the first word to the last word in the text document and to pass each word to the next module or halt the word checking process if there are no more words in the text document. If there is a word in the text document, the third module, Word Processing module, checks each word against root word dictionary lookup (e.g. *berat*, *pasukan*, *menteri*), proper nouns (e.g. *Kedah*, *Azlan*, *Proton*), abbreviation (e.g. *MITI, Jan* and *OKU*) and English words (e.g. *product*, *personnel* and *professional*) that have similar morpheme with affixation words so that text stemming rules do not stem these words. Another function of Word Processing module is to ensure only affixation words are delivered to text stemming module. The fourth module, Dictionary-based Text stemming module, stems each word that has conflicting morphological or exceptional rules in Malay language using derivative dictionary lookups which contains derived words and their respective root words (e.g. [*perubahan*, *ubah*], [*perompak, rompak*]). For instance, the word *menyanyi* and *menyapu* have two different morphological rules that may lead to stemming errors if not treated accordingly. The root word of the word *menyanyi* and *menyapu* are *nyany*i and *sapu* (not *nyapu*), respectively. Therefore, Dictionary-based Text stemming module, is to address these conflicting morphological rules or exceptional rules in the Malay language whereby sole dependence on ruled-based stemming will lead to various stemming errors. After the Dictionary-based Text stemming module has addressed possible stemming errors, the fifth module, Rule-based Text stemming module, stems prefixation, suffixation and confixation words to their respective root words using affixes removal method. Finally, the sixth module, Output module, will display the results of stemming process.

5 Experiment Results and Discussions

To evaluate the proposed text stemmer, 500 Malay online articles containing 112,785 word occurrences or 20,853 unique words and 114 chapters of Malay translations of Holy Qur'an containing 144,081 word occurrences or 11,387 unique words were used as testing datasets. The number of unique words in both testing datasets are imbalanced with respect to the number of total words despites of total number of words in 114 chapters of Malay translations of Holy Qur'an is more than 500 Malay online news articles. The main reasons are due to the numbers of word repetitions appeared in the Qur'an such as *hari* (day) [365 times], *pahala* (rewards) [117 times] and *akhirat* (hereafter) [115 times] whereas online news articles contain various categories: politics, economy, sports and world news that lead to more unique word occurrences in the text documents. Two sets of experiments were conducted to evaluate the stemming accuracy of the proposed text stemmer.

The first experiment showed that the performance of the proposed text stemmer achieved a 98.4% stemming accuracy on affixation words. Three main factors contributed to stemming accuracy are due to i.e., combinations of two words (e.g. *telahmelawan*), misspelled words (e.g. *ganguan*) and errors in character encoding (e.g. *noneperompak*). These stemming errors were not considered during the development of the proposed text stemmer, i.e., word tokenization (*telahmelawan* → *telah melawan*), word spell checker (*ganguan* → *gangguan*) and character encoding conversion from Unicode to Plain Text (*noneperompak* → *perompak*). Similarly, these stemming errors have not been considered by past researchers. In short, there was no evidence of affixation stemming errors in the first experiment as discussed in Section 4.

On the other hand, the second experiment showed that the performance of the proposed text stemmer on affixation words achieved a 97.5% stemming accuracy. Three main factors contributed to these stemming errors due to there is no stemming rules for removing affixes special characters (*al-*), affixed compounding words (*ketidakadilan*) affixed infixation words (*kesinambungan*) and words with special characters (*al-'araf*). These types of stemming errors also have not been considered by past researchers.

Our experimental results showed that the proposed text stemmer can stem prefixation, suffixation, confixation, and infixation words to their respective root words using rule-based approaches of 112 prefixation stemming rules, 97 suffixation stemming rules, 420 confixation stemming rules and dictionary-based approach of infixation stemming rule. Furthermore, the proposed text stemmer also uses root word dictionary lookup (e.g. [*berat*],[*makan*],[*kesan*]) and derivative dictionary lookup (e.g. [*perubahan*, *ubah*], [*perompak, rompak*]) containing only 10,527 word entries to address possible word-based and rule-based stemming errors. In contrast, current Malay stemmers with the stemming accuracy of 99.9% [14], 99.8% [7], 97.4% [16], 90% [11], 90.25% [10], use the root word dictionary from SISDOM98, which has 22,429 root words, to address only word-based stemming errors [5,7,11-12,16]. Even though some current text stemmers had high stemming accuracy, these stemmers were evaluated against minimum numbers of unique words (< 5,000 words) which not represent majority of affixation word distributions. It is important to note that most of the current text stemmers suffered from stemming errors due to generalizing affixation stemming rules to stem affixation words to their respective root words without considering the combinations of affixes, clitics, and particles that may need to be removed from the affixation words including special variations and exceptions for specific prefixes and confixes [5,7,9-16]. Moreover, the current text stemmers are highly dependent on a root word dictionary lookup [6], which could lead to understemming or overstemming errors as discussed in Section 4. Therefore, it has been suggested that root word and derivative dictionaries should be used to make affixation text stemming more effective and efficient in addressing these possible stemming errors. Furthermore, the word entries in a dictionary are not always useful as some root words are not affected by text stemming rules. The combination of affixation stemming and dictionaries lookup in the proposed text stemmer reduces possible affixation stemming errors due to root words with similar morphemes to affixes in affixation words (word-based stemming errors) and text stemming complexities to stem affixation words (rule-based stemming errors).

6 Conclusion

This paper discussed a proposed text stemmer that addresses possible stemming errors that could happen in the text stemming process. The proposed text stemmer eliminates four different types of possible stemming errors due to root words with similar morphemes to affixes in affixation words and the conflicting morphological rules in the Malay language. To address these stemming errors, the combination of ruled-based stemming method and dictionary lookup methods have been developed for achieving toward an error-free text stemming. Based on our experimental results, it can be concluded that the proposed text stemmer improves stemming accuracy in prefixation, suffixation, confixation and infixation text stemming. Our future work will focus on elevating the proposed text stemmer to include word tokenization and misspelled word checker to ensure that the input words are spelled correctly to avoid stemming errors.

References

1. Singh, J., & Gupta, V.: A systematic review of text stemming techniques. Artificial Intelligence Review, 48(2), 157-217 (2017).
2. Alfred, R., Ren, L. J., and Obit, J. H.: Assessing Factors that Influence the Performances of Automated Topic Selection for Malay Articles. International Conference on Soft Computing in Data Science, 300-309. Springer, Singapore (2016).
3. Willett, P. 2006. The Porter stemming algorithm: then and now. Program, 40(3), 219-223.
4. Hassan, A.: Morfologi (Vol. 13). PTS Professional (2006).
5. Ahmad, F., Yusoff, M., and Sembok, T. M.T: Experiments with a Stemming Algorithm for Malay Words. Journal of the American Society for Information Science, 47(12), 909-918 (1996)
6. Alfred, R., Leong, L. C., On, C. K., and Anthony, P.: A Literature Review and Discussion of Malay Rule-Based Affix Elimination Algorithms. The 8th International Conference on Knowledge Management in Organizations, 285-297. Springer, Netherlands (2014).
7. Darwis, S. A., Abdullah, R., and Idris, N.: Exhaustive Affix Stripping and A Malay Word Register to Solve Stemming Errors and Ambiguity Problem in Malay Stemmers. Malaysian Journal of Computer Science (2012).
8. Lovins, J. B.: Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory (1968).
9. Othman, A.: Pengakar Perkataan Melayu untuk Sistem Capaian Dokumen, MSc Thesis. Universiti Kebangsaan Malaysia. Bangi (1993).
10. Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., and Dhalila, M. S. S.: Simple Rules Malay Stemmer. The International Conference on Informatics and Applications (ICIA2012), The Society of Digital Information and Wireless Communication, 28-35 (2012).
11. Idris, N., and Syed, S. M. F. D.: Stemming for Term Conflation in Malay Texts. International Conference on Artificial Intelligence (2001).
12. Yasukawa, M., Lim, H. T., and Yokoo, H.: Stemming Malay Text and Its Application in Automatic Text Categorization. IEICE transactions on information and systems, 92(12), 2351-2359 (2009).
13. Abdullah, M. T., Ahmad, F., Mahmod, R., and Sembok, T. M. T.: Rules frequency order stemmer for Malay language. IJCSNS International Journal of Computer Science and Network Security, 9(2), 433-438 (2009).
14. Leong, L. C., Basri, S., and Alfred, R.: Enhancing Malay Stemming Algorithm with Background Knowledge. PRICAI 2012: Trends in Artificial Intelligence, 753-758. Springer, Berlin Heidelberg (2012).
15. Sankupellay, M., and Valliappan, S.: Malay Language Stemmer. Sunway Academic Journal, 3, 147-153 (2006).
16. Lee, J., Othman, R. M., and Mohamad, N. Z. 2013. Syllable-based Malay word stemmer. Computers & Informatics (ISCI), 2013 IEEE Symposium, 7-11. IEEE (2013).

17. Sembok, T. M. T., Yussoff, M., and Ahmad, F.: A Malay Stemming Algorithm for Information Retrieval. Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing, Vol. 5, 2-1 (1994).

Contactless Palm Vein ROI Extraction using Convex Hull Algorithm

Wee Lorn Jhinn, Michael Goh Kah Ong✉, Lau Siong Hoe, Tee Connie

Faculty of Information and Science Technology, Multimedia University, Melaka
michael.goh@mmu.edu.my

Abstract. In recent years, increased social concern towards hygienic biometric technology has led to a high demand for contactless palm vein biometric. Nonetheless, there are a number of challenges to be addressed in this technology. Among the most important challenges in the hand rotation issue that is caused inadvertently by unrestricted hand posture. In spite of the existing palm ROI region methods, the inadequacies of handling large rotations have never been accounted. In this paper, a rotation-invariant palm ROI detection method is proposed to handle a hand rotation of up to 360º and thus, providing a high flexibility for hand placement on the sensor. Experiments on the benchmark database validate the effectiveness of the proposed contactless palm vein approach.

Keywords: contactless, palm vein biometric, ROI, rotation invariant, detection

1 Introduction

For years, authentication methods that applied PIN, alphanumeric password, or smartcard have been ubiquitous in daily life. Nonetheless, the emergence of biometric technology has given a fresh impetus as an alternative solution for verifying one's identity nowadays. Knowing that biometrics exploit human's physiological or behavioral traits as "password" ingeniously for authentication, the hand biometric modality has lit up a great interest among the other biometrics due to the convenience of use in practical applications [1].

The hand vein biometric can be generally classified into multiple modalities known as finger vein, palm veins, dorsal veins, and wrist vein respectively. Naturally, the palm vein biometric has caught the eyes of the researchers and industrials due to an immense structure of vein pattern and thus, contributing to a greater representation area of ROI for recognition. Not to mention that the palm area is unobstructed from skin color and hair, a clearer vein pattern structure can be further obtained as compared to the other hand vein biometrics. Besides that, a specified near infrared (NIR) spectrum of between 690nm~900nm can reveal the vein vessels that are lying beneath the hand skin. Furthermore, this range of NIR is said to be sufficient for revealing the vein vessel structure without causing any healthy harming issue towards human body [2].

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_3

Back in year 2009, the outbreak of SARS disease in some Asian countries had raised the concern of the populace towards the public hygienic issue as the physical interaction between subject and biometric capture device was required [3]. Subsequent to that, this had impelled the development of contactless palm vein biometric.

On the basis of being contactless, the existing contactless palm vein systems relied on some assistive mechanisms for fixing the hand positions. These biometric capture devices are attached with a pegs support to restrict the hand posture regularly. Consequently, a novel rotation-invariant hand detection method is presented in this paper to offer a peg-free environment, where users can position their hand freely above the capture sensor within a visible distance, during the verification process.

2 Related Works

A number of palm ROI detection algorithm are discussed in the literature. Wang et al. [4] had proposed a novel ROI extraction approach for palm print and palm vein simultaneously. Nonetheless, they emphasized on the necessity of a fixed and restricted hand position coordinate system set up during the scanning process. In the meantime, Michael et al. [5] had put forward another novel palm ROI extraction method that can handle a maximum rotation of up to 30º under a contactless scenario. Although Ouyang et al. [6] claimed that their proposed rotation detection mechanism, which further contained of correction mechanism by using neural network, is able to perform under any rotation, the illustrated empirical results contained of only the hand rotation scenario within 45º. On the other hand, the method proposed by El-Sallam et al. [7] had adopted a simple rotation correction method in order to handle the hand pose contained of variety fingers' orientation. Inspired by that, Kang et al [8] had proposed a method that can tolerate a maximum rotation of up to 60º.

3 Proposed Work

In this work, a rotation invariant ROI detection algorithm for contactless palm vein biometric system is proposed. The overall framework of the proposed approach is illustrated in Fig. 1.

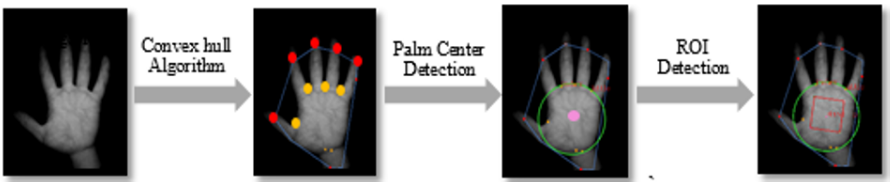


Fig. 1. The proposed rotation-invariant ROI detection framework system.

The input hand image is acquired by an own designed NIR acquisition device. After that, some pre-processing methods are adopted to segment the hand boundary to reduce the computation burden. The process is followed by further applying the convex hull algorithm in order to distinguish the fingertips point and hand valley point from the hand image differently. With the presence of the fingertips and hand valleys, the structure of palm ROI commences through the proposed technique.

3.1 Palm ROI Detection

In this paper, the Sklansky's convex hull algorithm [9] that employs the notion of convex polygon principle is adopted. Besides, OTSU thresholding method [10] and Suzuki contour detection method [11] have been applied in advance to binarized the hand image and extract the hand boundary of the hand, known as the hand contour. The relay of the Sklansky's convex hull algorithm begins with a given set of hand contour points, denoted as $S_{i=1,2,3,4\ldots n}$ where $n \in$ hand contour points. Fig. 2 illustrates the steps for the selection of H_1, H_2 , and H_3 in a row among some synthetically selected scatter points on the hand contour.

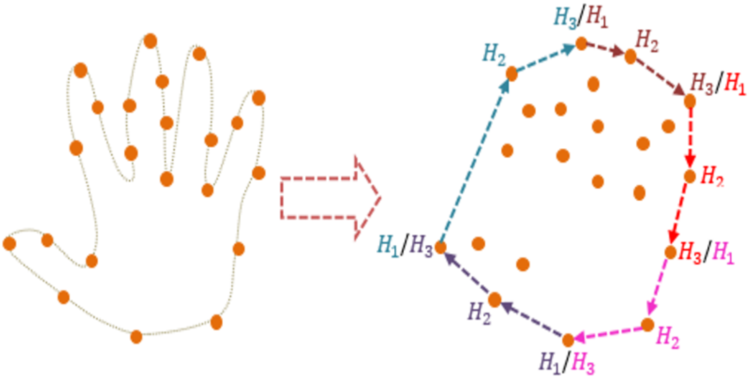


Fig. 2. The moving sequence of selected convex hulls.

The leftmost scatter point with the smallest x-coordinate is first selected as the starting point, which is potentially a convex hull, to form the convex polygon. Once the starting point denoted as H_1, is determined, the second point, H_2, is chosen randomly among the points close to H_1. The third point, H_3, is chosen randomly among S that is close to H_2. These three points have to be determined in order to form a virtual forward route. Eventually, if the angle between H_1 and H_3 is less than 180°, H_1, H_2 and H_3 are treated as a set of valid convex hull points.

On the contrary, if the angle between H_1 and H_3 is more than 180°, H_2 becomes a concave point. If the angle is concave, H_2 is treated as the hand valley point and the test is reversed back to start with H_1 and another two new possible convex hull points: $H_{1\_new}, H_{2\_new}, H_{3\_new}$. The angle between H_{1_new} and H_{3_new} has to be less than 180°. Fig. 3 illustrates the convex hull (i.e. fingertip points) and concave point (i.e. valley points) found on the hand contour.

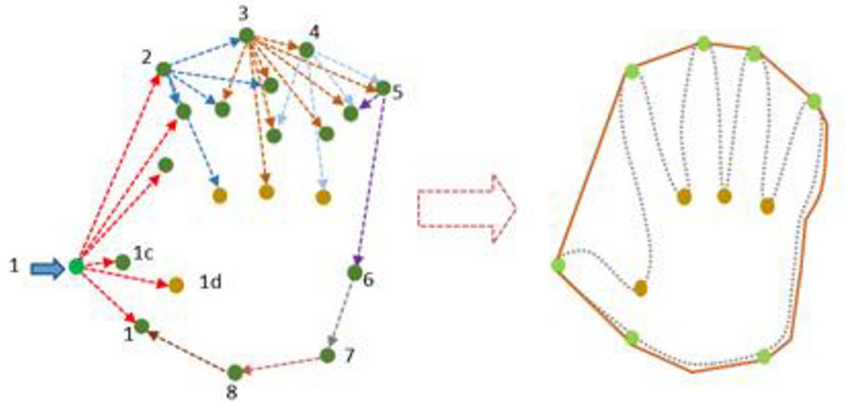


Fig. 3. The final convex polygon formed to enclose the entire hand.

3.2 Finger Tips and Valley Points Verification

With a list of the discovered convex hull and concave points, a two level verification process inspired by Goh et al. [5] valley searching is applied. In general, this verification process is desired to discard any incorrect labelled fingertips and valley points by creating a formation of a circle with a defined radius from each convex hull point. Hence, every convex hull point acts as the center point of its own neighbourhood circle. In the first checking stage, four basis points are used as the foundation for completing the circle formation of a convex hull as shown in Fig. 4 (a). The four basis points should have at least one point that is not lying within the enclosed contour. Otherwise the checking mechanism will instantly discard the contour object.

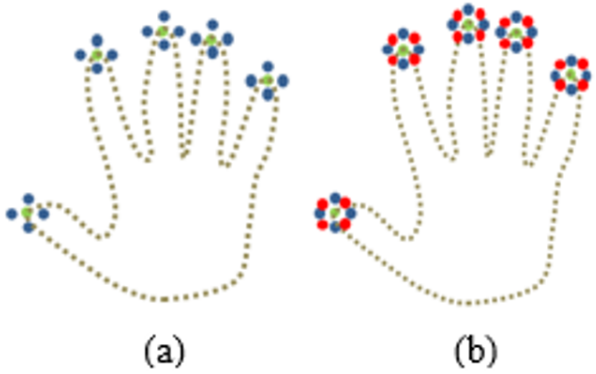


Fig. 4. (a) The circle formation on fingertips (b) The second stage for verifying correct fingertips.

If the convex hull has qualified in the first level, the circle edge is filled to consummate a complete circle. Note that the circle is formed synthetically in consideration of rotation invariant issue. Besides, it is easier to count the edge points in a circle. The total edge points of the circle are used as the core for final fingertips verification. Assuming that a complete circle shape has been established successfully in the second checking stage, any convex hulls, which are the center point of its own circle, should have 8 basis points apart from it concurrently. The convex hulls that are located on the peak of fingertips as shown in Fig. 4 (b) should have at least two edge points but not more than three lying in the enclosed hand contour. After all fingertips and valley points have been verified, our previously proposed ROI detection technique [13] will be applied to crop the correct palm ROI region through constructing a ROI square box.

4 Experimental Setup and Assessment

In this experiment, the 940nm CASIA benchmark database and an in-house database are employed for the assessment of the proposed palm ROI detection method in 12 different angles, respectively. A laptop with Intel® Core i7-4500U (2.39GHz) and 4GB RAM was used for this evaluation.

4.1 Technique Performance Evaluation

There are 100 subjects with 6 samples per subject in CASIA and 50 subjects with 10 samples per subject for the in-house database. Therefore, there is a total of 600 and 500 sample images for CASIA and in-house database in 12 different angles, respectively. The palm ROI cropping success rate of 12 different angles for the left and right hands for CASIA and in-house database are shown in Table 1 and Table 2. In spite of the rotation of image on any angles, the ratio of the hand image size is retained to preserve a fine image quality for the purpose of applying a precise ROI detection.

Table 1. CASIA correct ROI cropping rate

Angles	Left Hand	Right Hand
30°	95.83% (575)	98.00% (588)
60°	95.67% (574)	98.33% (590)
90°	89.67% (538)	93.33% (560)
120°	95.00% (570)	96.83% (581)
150°	95.50% (573)	97.83% (587)
180°	94.00% (540)	94.33% (566)
210°	95.50% (573)	96.17% (577)
240°	92.68% (556)	95.33% (572)
270°	90.33% (542)	93.67% (562)
300°	95.33% (572)	96.67% (580)
330°	96.33% (578)	97.50% (585)
360°/0°	95.17% (571)	97.67% (586)

Table 2. In-house database correct ROI cropping rate

Angles	Left Hand	Right Hand
30°	98.00% (490)	99.00% (495)
60°	98.00% (490)	99.20% (496)
90°	94.40% (472)	97.00% (485)
120°	97.40% (487)	98.00% (490)
150°	98.00% (490)	97.60% (488)
180°	99.00% (492)	98.40% (492)
210°	96.00% (480)	97.60% (488)
240°	96.60% (483)	96.00% (480)
270°	95.00% (475)	97.20% (486)
300°	99.20% (496)	99.00% (495)
330°	99.40% (497)	99.80% (499)
360°/0°	99.20% (496)	99.40% (497)

Although the left hand variations tend to be higher than the right hand for both databases, there are two general factors: 1) hand posture and 2) angles problem that are distorting the accuracy of ROI detection. It is observed that the correct ROI rate for both the left and right hands of two the databases are shown to have a more conspicuous result than 90°, 120°, and 270°, where 90° and 270° are known to be appertained to the category of Quadrantal Angles (QA). Indeed, the occurrence of this correct ROI cropping rate in QA is partially affected by a specific hand posture that results in inaccurate fingers labelling, where the index finger is often treated as the longest fingertip compared to the middle or ring finger. Fig. 5 depicts the mentioned specific hand posture from CASIA database that will distort the fingers labelling for both hands and consequently derive a wrong palm ROI.

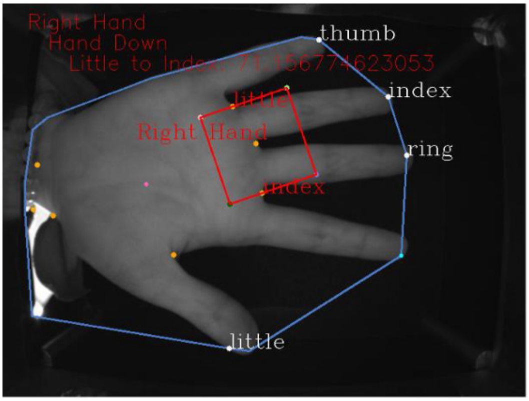


Fig. 5. The specific hand posture that resulting in wrong ROI detection

Moreover, aside of the mentioned hand posture issue as the first factor that affects the correct ROI cropping rate, the second factor that degrades the ROI cropping rate comes from an angle problem named as Quadrantal Angles Problem (QAP). In Fig. 6, it can be observed that the detection of the longest fingertips and fingertips label are correct. However, the ROI formulation persists to apply the inappropriate formula due to that extremely minimal difference of the predefined angles for determining the hand direction. The QAP issue is mainly caused by an inappropriate ROI application within the QA region and thus, deriving a wrong domain as the desired ROI. To be precise, the examination of the hand side to be facing up or down turns out to be ambiguous in QA.

Logically, the ROI formulation was applied precisely as it fulfills the requirement of hand facing up and the angle difference between the little valleys to index valley is 250.99°. Nonetheless, the hand was determined to be facing up as it entered the Q1 angle region from the palm center to the longest fingertip, which is logically correct, yet this has to be determined as facing down under this circumstance in order to apply the correct ROI formula. Thus, there is a minor angle difference of ± 15° through exhaustive testing in some CASIA hand images that will possibly lead to a wrong hand direction determination.

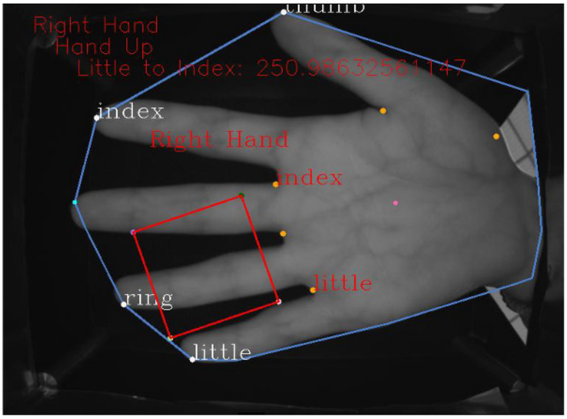
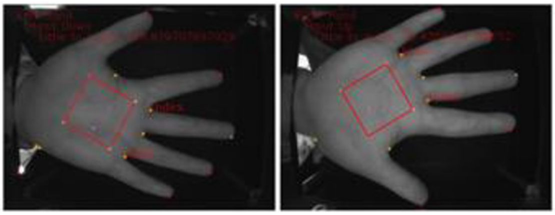


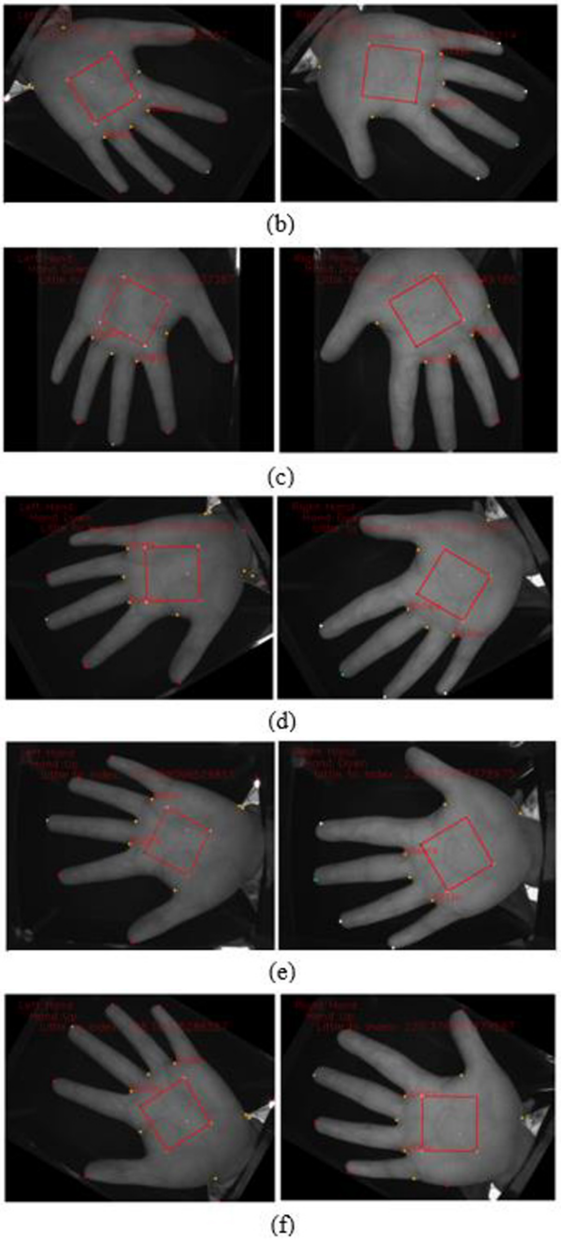
Fig. 6. The wrong ROI detection in 270º

It can be seen that the ROI cropping success rate drops when it comes to 90° and 270° for both databases. The CASIA left hand correct ROI cropping rate for these two angles are 89.67% and 90.33%, while the right hand ROI cropping success rate are 93.33% and 93.67%, respectively. On the other hand, the in-house database yields a better ROI cropping result, which is 94.40% in 90° and 95% in 270° for left hand while 97% in 90° and 270° for right hand. It can be further observed that the right hand provides a better ROI cropping success rate as compared to the left hand among these quadrant angles. This scene is interpreted as the variation of left hand is higher as compared to the right hand, especially when the image is further rotated and it becomes trickier to locate a correct ROI under such high variation. Meanwhile, the left hand images of the database are also interestingly facing the same situation, where the left hand images persist to contain a higher variation than the right hand images.

In spite of the ROI cropping success rate in QA, the success rate for the other angles are stable. The peak ROI cropping success rate for CASIA left and right hand are located at 30° and 60° that yield the correct ROI cropping rate of 95.83% and 98%, respectively. On the other hand, the peak result of correct ROI cropping for both hands in the in-house database takes place at 330° simultaneously with the result of 99.40% and 99.80% for left and right hand individually. More importantly, this has shown that the proposed ROI cropping technique is able to cope with most rotations and perform relatively stable ROI detection under a normal ratio quality image except the QA region. However, notice that the ROI cropping rate at 240° for CASIA both hands did not exceed the own hands success rate threshold. In the view of rotation, this issue occurred due to 240° is very close to 270° and thus, triggered the QAP. Fig. 7 depicts the final ROI detection result on some CASIA left and right hand images after applying the proposed ROI method based on a few unusual angles.



(a)



(b)

(c)

(d)

(e)

(f)

Fig. 7. The ROI detection result on (a) 90° (b) 120° (c) 180° (d) 240° (e) 270° (f) 300° of CASIA left hand (left side) and right hand (right side) images.

5 Conclusion

In this work, a palm ROI rotation invariant algorithm is proposed. The proposed approach works well for large hand rotation problem. Moreover, the addition of verifying proper fingertips and valley points has further increased the stability of the ROI detection. Nonetheless, the proposed ROI rotation invariant algorithm is still lacking for individuals with middle or ring finger deformity. Although the proposed method works favorable for most cases, we believe that further assessment have to be conducted in order to improve the robustness of the ROI detection method in the future.

Acknowledgement. The authors thankfully acknowledge the Chinese Academy of Sciences for providing CASIA-MS-PalmprintV1 database used in this work.

References

1. Wang, L. et al. "Infrared Imaging Of Hand Vein Patterns For Biometric Purposes". *IET Computer Vision*, vol 1, no. 3, 2007, pp. 113-122. *Institution Of Engineering And Technology (IET)*, doi:10.1049/iet-cvi:20070009.
2. Kim, J.G. et al. "Extinction Coefficients Of Hemoglobin For Near-Infrared Spectroscopy Of Tissue". *IEEE Engineering In Medicine And Biology Magazine*, vol 24, no. 2, 2005, pp. 118-121. *Institute Of Electrical And Electronics Engineers (IEEE)*, doi:10.1109/memb.2005.1411359.
3. Ong Michael, Goh Kah et al. "A Contactless Biometric System Using Palm Print And Palm Vein Features". *Advanced Biometric Technologies*, 2011. *Intech*, doi:10.5772/19337. Accessed 21 Mar 2018.
4. Wang, Jian-Gang et al. "Person Recognition By Fusing Palmprint And Palm Vein Images Based On “Laplacianpalm” Representation". *Pattern Recognition*, vol 41, no. 5, 2008, pp. 1514-1527. *Elsevier BV*, doi:10.1016/j.patcog.2007.10.021.
5. Michael, Goh Kah Ong et al. "Robust Palm Print And Knuckle Print Recognition System Using A Contactless Approach". *2010 5Th IEEE Conference On Industrial Electronics And Applications*, 2010. *IEEE*, doi:10.1109/iciea.2010.5516864. Accessed 21 Mar 2018.
6. Ouyang, Chen-Sen et al. "An Improved Neural-Network-Based Palm Biometric System With Rotation Detection Mechanism". *2010 International Conference On Machine Learning And Cybernetics*, 2010. *IEEE*, doi:10.1109/icmlc.2010.5580754. Accessed 21 Mar 2018.
7. El-Sallam, A. et al. "Robust Pose Invariant Shape-Based Hand Recognition". *2011 6Th IEEE Conference On Industrial Electronics And Applications*, 2011. *IEEE*, doi:10.1109/iciea.2011.5975595. Accessed 21 Mar 2018.
8. Kang, Wenxiong, and Qiuxia Wu. "Contactless Palm Vein Recognition Using A Mutual Foreground-Based Local Binary Pattern". *IEEE Transactions On Information Forensics And Security*, vol 9, no. 11, 2014, pp. 1974-1985. *Institute Of Electrical And Electronics Engineers (IEEE)*, doi:10.1109/tifs.2014.2361020. Accessed 21 Mar 2018.
9. Sklansky, Jack. "Finding The Convex Hull Of A Simple Polygon". *Pattern Recognition Letters*, vol 1, no. 2, 1982, pp. 79-83. *Elsevier BV*, doi:10.1016/0167-8655(82)90016-2. Accessed 21 Mar 2018.

10. Otsu, Nobuyuki. "A Threshold Selection Method From Gray-Level Histograms". *IEEE Transactions On Systems, Man, And Cybernetics*, vol 9, no. 1, 1979, pp. 62-66. *Institute Of Electrical And Electronics Engineers (IEEE)*, doi:10.1109/tsmc.1979.4310076. Accessed 21 Mar 2018.
11. Suzuki, Satoshi, and KeiichiA be. "Topological Structural Analysis Of Digitized Binary Images By Border Following". *Computer Vision, Graphics, And Image Processing*, vol 30, no. 1, 1985, pp. 32-46. *Elsevier BV*, doi:10.1016/0734-189x(85)90016-7. Accessed 21 Mar 2018.
12. Skyum, Sven. "A Simple Algorithm For Computing The Smallest Enclosing Circle". *Information Processing Letters*, vol 37, no. 3, 1991, pp. 121-125. *Elsevier BV*, doi:10.1016/0020-0190(91)90030-l. Accessed 21 Mar 2018.
13. Jhinn, Wee Lorn et al. "A Contactless Rotation-Invariant Palm Vein Recognition System". *Advanced Science Letters*, vol 24, no. 2, 2018, pp. 1143-1148. *American Scientific Publishers*, doi:10.1166/asl.2018.10704.

A Robust Abnormal Behavior Detection Method Using Convolutional Neural Network

Nian Chi Tay[1], Tee Connie[1], Thian Song Ong[1], Kah Ong Michael Goh[1] and Pin Shen Teh[2]

[1] Multimedia University, Malacca, Malaysia
[2] University of Manchester, Manchester, UK
nianchi.tay95@gmail.com

Abstract. A behavior is considered abnormal when it is seen as unusual under certain contexts. The definition for abnormal behavior varies depending on situations. For example, people running in a field is considered normal but is deemed abnormal if it takes place in a mall. Similarly, loitering in the alleys, fighting or pushing each other in public areas are considered abnormal under specific circumstances. Abnormal behavior detection is crucial due to the increasing crime rate in the society. If an abnormal behavior can be detected earlier, tragedies can be avoided. In recent years, deep learning has been widely applied in the computer vision field and has acquired great success for human detection. In particular, Convolutional Neural Network (CNN) has shown to have achieved state-of-the-art performance in human detection. In this paper, a CNN-based abnormal behavior detection method is presented. The proposed approach automatically learns the most discriminative characteristics pertaining to human behavior from a large pool of videos containing normal and abnormal behaviors. Since the interpretation for abnormal behavior varies across contexts, extensive experiments have been carried out to assess various conditions and scopes including crowd and single person behavior detection and recognition. The proposed method represents an end-to-end solution to deal with abnormal behavior under different conditions including variations in background, number of subjects (individual, two persons or crowd), and a range of diverse unusual human activities. Experiments on five benchmark datasets validate the performance of the proposed approach.

Keywords: Abnormal behavior detection, Convolutional Neural Network, Deep learning.

1 Introduction

There is a pressing need for tightened security due to the increasing crime rate in the society. Every now and then, there are headlines and news about crime cases such as robbery, personal attack, and terrorism. To deter criminal offenses and to ensure public safety, surveillance devices like CCTV cameras have been installed in public places such as banks, schools, shops and subway stations. However, it is impractical for

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_4

human to effectively monitor the cameras twenty-four hours a day, seven days a week. This is where computer vision technology comes in. Today's modern surveillance system not only aims to monitor and substitute the human eye, but also to carry out surveillance automatically and autonomously. The perception for abnormal behavior differs on situations. A behavior is said to be abnormal if the behavior is different from one's neighbors [1]. For example, the running action is considered normal in a field but is considered abnormal if it happens in a shopping mall. If an abnormal behavior can be detected early by the surveillance system, many tragedies can be prevented from happening.

This paper proposes a deep learning approach for abnormal behavior detection. Deep learning is inspired by neural network which contains a deep structure to learn useful features and representations directly from the data. A typical neural network is made up of an input layer, several hidden layers and an output layer. A deep network, on the other hand, consists of a large network comprising of many layered networks [2]. Convolutional Neural Network (CNN) is one of the popular networks in deep learning.

In this work, we present a CNN-based method for abnormal behavior detection. The method automatically learns the characteristics concerning a wide range of abnormal behaviors. We also analyze the performance of the proposed method using various subjects such as individual, two persons, and crowd behaviors involving different background settings. Such diverse analysis has not been studied before.

This paper is organized into four sections. Section 2 discusses the related works on abnormal behavior detection. Section 3 introduces our proposed CNN framework. Section 4 discusses the experiment and results obtained. Section 5 presents the conclusion and future work.

2 Related Works

There are various methods and techniques used for human abnormal behavior detection in surveillance system. In this paper, we focus on the most crucial components in CNN: training and learning of data. The data are fed to the network to learn useful features about the data in order to perform recognition. The existing approaches can generally be categorized into three broad categories.

The first category is supervised learning. This is a type of learning network whereby the labels of normal and abnormal behaviors are given beforehand correspond to the situations. The network takes the input features and also the labels for training [3] - [6]. If the label of the test sample matches the training sample that contains normal behavior, it is classified as normal behavior, whereas if not, then is classified as abnormal behavior. The second category is unsupervised learning. This is a type of learning whereby the network clusters the data without any labels [7] - [10]. In order to cluster the data into abnormal or normal behavior, certain statistical properties and methods are needed. The data that have similar features are clustered in the same group whereas isolated clusters are defined as anomalies, which represent the abnormal behaviors. The last category is semi-supervised learning. This is a type of learning whereby it requires a mixture of labeled and unlabeled data [11]-[14]. This approach inherits the advantages and disadvantages of both methods which will be discussed in the later part of the paper.

For supervised learning approach, Ko et al. [3] proposed deep convolutional framework. The input image is first fed into the CNN and applied with Kalman filter. Next, the output vector is transferred to Long Short Term Memory (LSTM) network to perform the behavior classification. Kuklyte [4] implemented Motion Boundary Histogram (MBH) to segment spatio-temporal regions and SVM method to classify the data. Radial Basis Function (RBF) kernel is used to tackle the noise. Nater et al. [5] applied tracker trees method which specified the actions at a higher level of trees. For instance, detection at the lowest level was to recognize human. Further levels upwards were to identify specific actions such as unusual behaviors. The authors used appearance based probabilistic tracking to identify images that were represented in different forms like segmented, rescaled, distance transformed, embedded and reconstructed. The work by Lv et al. [6] performed features matching using Pyramid Match Kernel algorithm. The input actions in human silhouette form were modeled as 2D human poses and represented using Action Net which is a graph model.

For unsupervised learning method, Choudhary et al. [7] proposed Probabilistic Latent Semantic Allocation (pLSA) to extract the spatio-temporal features from videos containing indoor corridor monitoring that are segmented using video epitomes. On the other hand, Hu et al. [8] applied Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) to learn the abnormal or normal features from MIT PLIA 1 dataset which contains domestic house chores that are filtered by One-Class Support Vector Machine (OCSVM) model. The work by Varadarajan et al. [9] also used pLSA to recognize patterns in the busy traffic scenes. The scenes were then segmented into regions with particular activities using the low-level features extracted. Zhang et al. [10] introduced a three-phase approach that first used HDP-HMM to create classifiers. In the second phase, abnormal events were identified by an ensemble learning algorithm. Lastly, abnormal behavior models were derived from normal behavior model to decrease the false positive rate, which is the wrongly classified outputs in abnormal activity samples.

For semi-supervised learning technique, Wang et al. [11] combined k-means algorithm and Posterior Probability (PPSVM) to detect the classes from an imbalanced data. This method classifies the data using probability distributions and not features. Zou et al. [12] presented a semi-supervised Expectation-Maximization algorithm and extracted the features using Gaussian-based appearance similarity model to form histograms. Jager et al. [13] employed a three-phase learning procedure. The image sequences were first encoded using hidden Markov models (CHMMs) before the learning steps. In the first phase, one-class learning was carried out. Next, regular sequence model (RSM) was applied to detect the outliers. Lastly, the unusual segments were employed to expand RSM to form an error sequence model (ESM) which was controlled by Bayesian Information Criterion (BIC). The work by Li et al. [14] presented a four-steps method to detect abnormality. First, samples were obtained using Dynamic time warping (DTW) clustering method. Next, the parameters in HMM were trained by iterative learning approach. Maximum a posteriori (MAP) technique was used to estimate the parameters of abnormal behaviors from normal behaviors. Lastly, topological HMM was built to classify the abnormal behaviors.

Supervised learning is the simplest approach as compared to its unsupervised and semi-supervised counterparts. However, it is not very practical to be implemented in real world. This is because there are too many types of abnormal behaviors in practice, and a large number of data is needed for the network to learn and perform well for different scenarios. The existing labeled abnormal data are also hard to find and are often costly. On the contrary, unsupervised learning utilizes the statistics learned from unlabeled data samples to cluster normal and abnormal behaviors. The cost of implementing unsupervised learning approach is low. However, it might not obtain high accuracy due to the fact that the labels are undefined and it depends on statistical approach to cluster the labels. Semi-supervised is said to be the hardest method as it is challenging to discover how to deal with the mixture of labeled and unlabeled data for training. But one of the advantages of semi-supervised learning is that it solves the problem of insufficient labeled data and the mixture of cheap unlabeled data can be used together for training.

3 Proposed Approach

In this section, we provide the detail for the proposed approach. The input images are converted from video sequences containing normal and abnormal behaviors such as walking, jogging, fighting, kicking and punching. The RGB images are selected manually using eye inspection and the images undergo a pre-processing stage by applying a 3x3 moving average filter to remove noises in the images, $y_{ij} = \sum_{k=-m}^{m} \sum_{l=-m}^{m} w_{kl} x_{i+k,j+l}$ where x_{ij} denotes the input image and i, j represent the number of pixels in the image. The image output is referred to as y_{ij}. A linear filter of size 3 x 3 is used where ($2m$ + 1) x ($2m$ + 1) with weights w_{kl} for $k, l = -m, \ldots, m$ and m equals to 1 [15].

The video frames are manually sampled from the video sequences. Some important information might be lost when sampling the frames from the video sequences. High concentration is needed when selecting the frames to form normal or abnormal behavior dataset as some abnormal behaviors only occur in the middle of the video. Actions in the rest of the frames are categorized as normal behavior. The images are stored in image datastore, and the labels are assigned manually (also known as supervised learning) to each training image. CNN consists of three main components: the input layer which contains the input image, the middle layers which are also known as the feature detection layers, and the final layer which is the classification layer. The images in different sizes (due to video sequences obtained from different datasets) are resized to 32x32 pixels for speedy training. The input image goes through middle layers that consist of three operations: convolution, pooling and Rectified Linear Unit (ReLU). This paper uses 6 layers that consist of 3 convolution layers, 2 fully connected layers and a softmax layer. The framework of our CNN is shown in Fig.1.

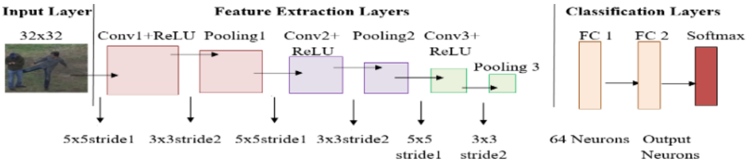


Fig. 1. The proposed CNN framework for abnormal behavior detection.

The convolution layer filters the input image and activates certain features of the image for example the edge, corner and texture information. The features are useful to detect the type of action being performed, and how compact the scene is (e.g. people pushing each other). The first convolutional layer has a number of 32 (5x5x3) filters. The third dimension refers to the colored-images input. A padding of 2 pixels is added symmetrically to ensure that image borders are taken into account. This step is important as it prevents the borders from being eliminated too early in the network. Next, the ReLU layer is added to map negative values to zero and to ensure there are only positive values. The ReLU layer allows faster training in the network. This is followed by a max pooling layer which has a 3x3 spatial pooling area with strides of 2 pixels. The size of the data is then down-sampled from 32x32 to 15x15. The three layers of convolutional, ReLU and pooling are repeated two times to complete the feature extraction layers. We avoid using too many pooling layers to prevent downsampling the data prematurely as some of the important features might be discarded too early.

After performing feature extraction, the network performs classification. There are basically two layers that form the final layers of the network for classification. The final layers consist of fully connected layers and a softmax layer. The first fully connected layer is made up of 64 output neurons from the input size of 32x32. A ReLU layer is added after that. Next, the second fully connected layer is used to output the number of signals which are the categories to be classified. For Experiment 1, the categories are abnormal and normal behaviors, whereas for Experiment 2, there are six categories include punching, kicking, pushing, hand-shaking, pointing, and hugging. Lastly, a softmax loss layer and a classification layer are used to calculate the probability of distribution for each category. The input layer, middle layers and final layers are combined together to form the complete network.

The first weights in the convolutional layer are initialized using normally distributed random numbers with 0.0001 as the standard deviation to decrease the loss when the learning of network takes place. This paper uses stochastic gradient descent with momentum (SGDM) to train the network. We tune the parameters inside the network to find out which features affect the outcome of the results. In this paper, the number of epochs is tuned from 10 to 100 with a step size of 10 and the initial learning rate is configured from 0.001 to 0.1. The number of epoch is a complete forward and backward passing of the training samples while the learning rate refers to the speed of finding the correct weights in the network. Deep learning often requires a large number of inputs to obtain the best accuracy. It also relies heavily on the computational resources and requires a high-performance GPU. The experiments in this paper are carried out using Matlab R2017b version on a workstation equipped with Intel® HD Graphics 5500 8GB CPU. A summary of the proposed CNN framework is shown in Table 1.

Table 1. Summary of CNN configuration.

Parameters	Conv1, Pool1, ReLU1	Conv2, Pool2, ReLU2	Conv3, Pool3, ReLU3
Conv. Filters	5x5	5x5	5x5
Conv. Stride	1	1	1
Conv. Padding	2	2	2
Max Pooling Filters	3x3	3x3	3x3
Max Pooling Stride	2	2	2
Kernels	32	32	32

4 Experiments and Results

4.1 Dataset Description

In this paper, five benchmark databases have been tested namely CMU Graphics Lab Motion Capture Database (CMU) [16], UT-Interaction dataset (UTI) [17], Peliculas Dataset (PEL) [18], Hockey Fighting Dataset (HOF) [19], and Web Dataset (WED) [20]. All datasets have different background settings such as indoor, game field, lawn, public places like pedestrian crossing and movie scenes.

The CMU dataset contains 11 videos with 6 normal and 5 abnormal behaviors. Normal behaviors include walking, hand-shaking, and jogging. Abnormal behaviors include resistant actions or violent gestures. For example, subject A pulls subject B by elbow but subject B resists; A pulls B by hand but B resists; A and B quarrel with angry hand gestures; A picks up a high stool and threatens to throw at B. There are a total of 2477 images, with 1209 positive images and 1268 negative images. There are 800 positive and negative images each for training, and 409 positive and 468 negative images for testing. The images are in RGB format in the size of 352x240 pixels, which are then resized to smaller pixels of 32x32 to shorten the training time.

The second dataset used is UTI dataset that consists of videos with 6 classes of human interactions. This includes 976 images of hand-shaking, 983 images of pointing, 904 images of hugging, 1027 images of pushing, 872 images of kicking and 847 images of punching. The dataset is taken on a lawn outdoor. 30 videos of abnormal behaviors and 24 videos of normal behaviors are selected. In this paper, we categorize pushing, kicking and punching as abnormal behaviors while hand-shaking, pointing and pushing as normal behaviors. There are a total number of 5609 images, 2706 positive images and 2903 negative images. This dataset is used to perform both binary and multi-class classifications. In the first part of the experiment, binary classification is carried out to identify normal and abnormal behaviors. 1800 positive and negative images are used for training, while 906 positive images and 1103 images for testing. The images are in RGB format in size of 276x236 pixels which are then resized to 32x32 pixels. In the second part of the experiment, multi-class classification is performed to categorize the images into six categories using 650 images for training and testing.

The third dataset used is the PEL dataset that consists of 368 images. There are 268 fighting images and 100 non-fighting images. The dataset consists of fighting scenes from movies. We categorize the fighting behavior as abnormal behavior and non-fighting behavior as normal behavior. There are 80 positive and negative images respectively for training, while 188 positive images and 20 images for testing. The images are in RGB format in size of 352x240 pixels which are resized to 32x32 pixels.

The fourth dataset used is the HOF dataset that consists of 1800 images with 900 positive and negative images respectively for training. As for the testing set, we use 600 images each as positive and negative images. The dataset is taken on real life hockey games when fighting against players happened. The positive images consist of fighting behaviors and negative images consist of normal archery images from the UCF Dataset [21]. The images are in RGB format in size of 360x288 pixels before resized to 32x32 pixels.

The fifth dataset used is the WED that consists of abnormal crowd behaviors like running in chaos in a public place. There are a total of 1280 images with 640 positive and negative images respectively. The training set is 450 for both positive and negative images, the testing set is 190 for positive and negative images respectively. The images are in RGB format in size of 320x240 pixels which are resized to 32x32 pixels.

4.2 Results and Discussions

The experiment is carried out in two parts. The first part (Experiment 1) is to classify the images into binary classes; either abnormal or normal behavior using CMU, UTI, PEL, HOF and WED datasets; while the second part (Experiment 2) is to classify the images into 6 categories (punching, kicking, pushing as abnormal behaviors; handshaking, pointing and hugging as normal behaviors) using UTI dataset. The experiments are split into two parts to evaluate the effect of the number of classes on the performance of the network. Some screenshots for Experiment 1 and Experiment 2 are shown in Fig. 2 containing both abnormal and normal behaviors. The last row in the figure presents the six categories of actions for Experiment 2.

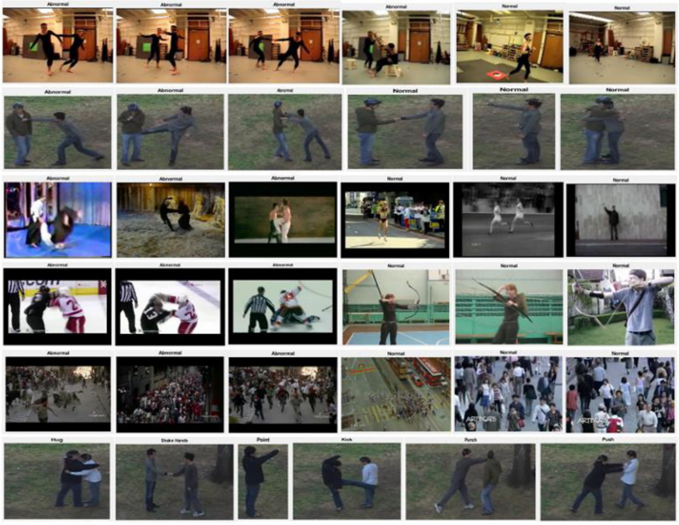


Fig. 2. Screenshots taken from datasets in rows: 1) CMU dataset, 2) UTI dataset, 3) PEL dataset, 4) HOF dataset, 5) WED dataset, 6) UTI dataset for Experiment 2.

Table 2 records the different number of epochs used for training with different learning rates in Experiment 1. It is shown that the proposed approach achieves high accuracy around 100% for all the datasets. A learning rate of 0.01 gives the highest accuracy for all the datasets. A learning rate of 0.001 can also achieve high accuracy, but the result is slightly lower for the UTI dataset and PEL dataset. The large number of behaviors in the first dataset makes it harder for the network to learn. The images in the PEL dataset are slightly blurred as compared to others. These may be the reasons of the slightly decreased performance. From the viewpoint of learning rate, low learning rate will cause slow convergence, overfitting and low accuracy. A learning rate of 0.1 is too fast for the network to learn the weights and this results in overshooting the global minimum. Apart from learning rate, the results also show that the higher the number of epochs, the better the accuracy. However, there is a risk for overfitting that results in a lower accuracy when it exceeds a certain number of epochs. The more epochs used, the more time-consuming the training is as shown in Table 3.

Table 2. Results obtained by using different learning rates in Experiment 1.

Learning Rate	Maximum Number of Epochs	Dataset Accuracy (%)				
		CMU	UTI	PEL	HOF	WED
0.001	10	99.66	54.90	90.38	58.50	89.21
	20	100.00	56.55	90.38	100.00	100.00
	30	100.00	57.74	90.38	100.00	100.00
	40	100.00	70.18	90.38	100.00	100.00
	50	100.00	99.15	90.38	100.00	100.00
	60	100.00	99.10	90.38	100.00	100.00
	70	100.00	99.70	90.38	100.00	100.00
	80	100.00	99.65	90.38	100.00	100.00
	90	100.00	99.70	90.38	100.00	100.00
	100	100.00	99.75	90.38	100.00	100.00
0.01	10	100.00	54.90	90.38	100.00	100.00
	20	100.00	99.60	90.38	100.00	100.00
	30	100.00	99.75	87.98	100.00	100.00
	40	100.00	99.55	100.00	100.00	100.00
	50	100.00	99.80	100.00	100.00	100.00
	60	100.00	99.80	100.00	100.00	100.00
	70	100.00	99.60	100.00	100.00	100.00
	80	100.00	99.70	100.00	100.00	100.00
	90	100.00	99.65	100.00	100.00	100.00
	100	100.00	99.80	100.00	100.00	100.00
0.1	10	46.64	54.90	9.62	50.00	0.00
	20	0.00	0.00	9.62	50.00	0.00
	30	0.00	54.90	0.00	50.00	0.00
	40	0.00	54.90	90.38	50.00	0.00
	50	0.00	0.00	9.62	50.00	50.00
	60	0.00	54.90	90.38	50.00	50.00
	70	0.00	0.00	9.62	50.00	50.00
	80	0.00	54.90	90.38	50.00	50.00
	90	0.00	54.90	90.38	50.00	0.00
	100	46.64	54.90	0.00	50.00	50.00

Table 3. Time taken to complete the training.

Dataset	Elapsed Time (s)									
	10	20	30	40	50	60	70	80	90	100
CMU	15.15	28.91	42.47	55.79	71.16	83.63	95.55	108.50	154.10	166.15
UTI	33.59	66.58	97.53	124.69	157.71	184.50	213.98	271.79	315.27	330.58
PEL	2.44	3.81	5.29	6.32	8.75	9.51	10.13	11.33	12.57	14.52
HOF	13.40	23.46	35.72	46.03	58.84	70.69	71.67	82.98	94.77	117.26
WED	9.98	16.71	23.96	32.07	40.26	47.24	55.86	62.96	70.15	79.13

For Experiment 2, the network is trained for multi-class classification. The results in Fig. 3 to Fig. 5 clearly show that the accuracy decreases when the learning rate approaches 0.1, and the network is unable to perform at all eventually. The accuracy starts to drop when it reaches a maximum number of epochs. Out of all the 6 behaviors, hand-shaking has a higher accuracy because this action does not have obstructed views as compared to other actions. The results obtained from Experiment 1 and Experiment 2 suggest that the number of categories in the training data does not affect the accuracy of the result as long as there are enough data provided for training. Table 4 provides a comparison of the proposed method with the state-of-the-art methods. It shows that the proposed method can achieve promising result for both Experiments 1 and 2 that consist of single-person behavior, two-person interactions and crowd behaviors. This demonstrates that the proposed approach is able to work well for abnormal behaviors across different settings.

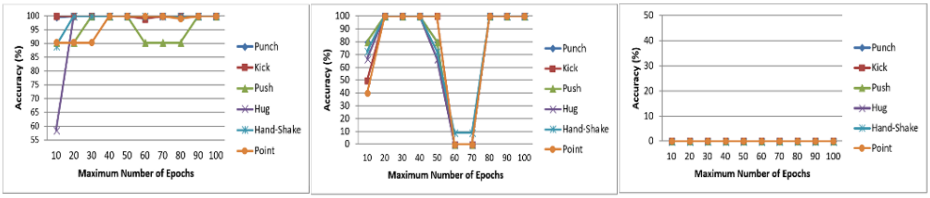


Fig. 3 - 5. Results obtained by using learning rate of 0.001, 0.01, and 0.1 in Experiment 2.

Table 4. Comparisons between the proposed method and the related works.

Authors	Methodology	Dataset Descriptions	Accuracy (%)
Ko and Sim [3]	CNN, Karman Filter and LSTM	UI-Interaction dataset	97%
Lv and Nevatia [6]	Pyramid Match Kernel algorithm	Action Net contains behaviors such as punch, kick, point and wave	80.6%
Zhang et al. [10]	Three-phase approach with HDP-HMM	CAVIAR sequences consist of walking, browsing and fighting behaviors	100% with 60% false alarm rate
Zou and Bhanu [12]	Gaussian-based appearance similarity model for feature extraction and Expectation-Maximization algorithm for classification	Human activities observed in a simulated camera network	100% with 4% false alarm rate
Jager et al. [13]	Three-phase learning procedure using CHMMs, RSM and ESM	Image sequence comprises up to 4000 frames	99.9% with 1.7% false positive rate
Proposed approach	CNN	5 datasets include UMI, UTI, HOF, WED and PEL containing behaviors such as kicking, fighting, punching, pushing, pulling etc.	Experiment 1: 100% Experiment 2: 100%

5 Conclusion and Future Work

This paper studies human abnormal behavior detection under different situations such as various background settings and number of subjects using convolutional neural network. Experiment results show that the proposed approach achieves favorable performance across different scenarios. Besides, the effects of different network configurations are examined. We demonstrate that the learning rate used for training should not be too high to avoid overshooting and not too low to prevent overfitting and low convergence of the network. The number of epochs should be tuned from a small value and gradually increased to achieve the highest accuracy.

In the future, we will explore abnormal behavior detection for single person, two persons and crowd under more diverse situations. This will help to design a more robust intelligent surveillance system that can tackle different types of practical situations.

References

1. Nian Chi Tay, P. S. Tay, S. W. Tay: "Deep Learning for Abnormal Behavior Detection", in Security and Authentication: Perspectives, Management and Challenges, Nova Science Publishers, United States (2018). (ISBN: 978-1-53612-942-7).
2. Cho, S., and Kang, H.:"Abnormal behavior detection using hybrid agents in crowded scenes," Pattern Recognition Letters, 44, 64-70. doi:10.1016/j.patrec.2013.11.017. (2014).
3. Ko, K., and Sim, K.: "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," Engineering Applications of Artificial Intelligence, 67, 226-234. doi:10.1016/j.engappai.2017.10.001. (2018).
4. Jogile Kuklyte: "Unusual event detection in real-world surveillance applications," Doctoral dissertation, Dublin City University. (2014).
5. F. Nater, H. Grabner, and L. Van Gool: "Exploiting simple hierarchies for unsupervised human behavior analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 13–18, pp. 2014–2021. (2010).
6. Fengjun. Lv and R. Nevatia: "Single view human action recognition using key pose matching and viterbi path searching," in Proc. IEEE Conf. Comput. Vision Pattern Recog., pp. 1–8. (2007).
7. A. Choudhary, M. Pal, S. Banerjee, and S. Chaudhury: "Unusual activity analysis using video epitomes and pLSA," in Proc. 6th Indian Conf. Comput. Vision, Graphics Image Process., pp. 390–397. (2008).
8. D. H. Hu, X. Zhang, J. Yin, V. W. Zheng, Q. Yang: "Abnormal activity recognition based on HDP-HMM models," in Proc. IJCAI, pp. 1715–1720. (2009).
9. J. Varadarajan and J. Odobez: "Topic models for scene analysis and abnormality detection," in Proc. IEEE 12th Int. Conf. Comput. Vision Workshops, Sep. 27–Oct. 4, pp. 1338–1345. (2009).
10. X. Zhang, H. Liu, Y. Gao, and D. H. Hu: "Detecting abnormal events via hierarchical Dirichlet processes," in Proc. 13th Pacific-Asia Conf. Knowledge Discovery Data Mining, Apr.27–30, pp. 278–289. (2009).
11. Wang, Y., Li, X., & Ding, X.:"Probabilistic framework of visual anomaly detection for unbalanced data". Neurocomputing, 201, 12-18. doi:10.1016/j.neucom.2016.03.038. (2016).
12. X. Zou and B. Bhanu: "Anomalous activity classification in the distributed camera network," in Proc. 15th IEEE Int. Conf. Image Process., pp. 781–784. (2008).
13. M. Jager, C. Knoll, and F. A. Hamprecht: "Weakly supervised learning of a classifier for unusual event detection," IEEE Trans. Image Process., vol. 17, no. 9, pp. 1700–1708, Sep. (2008).
14. H. Li, Z. Hu, Y. Wu, and F. Wu: "Behavior modeling and abnormality detection based on semi-supervised learning method," Ruan Jian Xue Bao/J. Software, vol. 18, pp. 527–537. (2007).

15. Glasbey, C. A., & Horgan, G. W.: "Chapter 3: Filters. In Image Analysis for the Biological Sciences." Wiley, United States. (1995).
16. "CMU Graphics Lab Motion Capture Database.", http://mocap.cs.cmu.edu/, last accessed 2018/1/2.
17. Ryoo, M.S. and Aggarwal, J.K.: "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)", http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.
18. "Peliculas Movies Fight Detection Dataset", http://academictorrents.com/details/70e0794e2292fc051a13f05ea6f5b6c16f3d3635/tech&hit=1&filelist=1, last accessed 2018/1/5.
19. E. Bermejo, O. Deniz, G. Bueno, R. Sukthankar: "Violence Detection in Video using Computer Vision Techniques", Proceedings of Computer Analysis of Images and Patterns. (2011).
20. "CRF Web Dataset", http://crcv.ucf.edu/projects/Abnormal_Crowd/#WebDataset, last accessed 2018/1/5.
21. "UCF101 Action Recognition Data Set", http://crcv.ucf.edu/data/UCF101.php, last accessed 2018/1/5.

Cryptanalysis of Improved and Provably Secure Three-Factor User Authentication Scheme for Wireless Sensor Networks

Jihyeon Ryu[1], Taeui Song[1], Jongho Moon[2], Hyoungshick Kim[3], and Dongho Won[3],*

[1]Department of Platform Software, Sungkyunkwan University
{jhryu, tusong}@security.re.kr
[2]Department of Electrical and Computer Engineering, Sungkyunkwan University
jhmoon@security.re.kr
[3]Department of Computer Engineering, Sungkyunkwan University
hyoung@skku.edu(H.K.); dhwon@security.re.kr(D.H.)

Abstract. Wireless sensor networks are applied in various areas like smart grid, environmental monitoring, health care, and security and surveillance. It applies to many fields, but as the utilization is higher, security becomes more important. Recently, the authentication scheme for the environment of wireless sensor network has also been studied. Wu et al. has announced a three-factor user authentication scheme claiming to be resistant to different types of attacks and maintain various security attributes. However, their proposal has several fatal vulnerabilities. First, it is vulnerable to the outsider attack. Second, it is exposed to user impersonation attack. Third, it does not satisfy user anonymity. Therefore, in this paper, we describe these vulnerabilities and prove Wu et al.'s scheme is unsafe.

Keywords: Wireless Sensor Network · Elliptic Curve Cryptosystem · Remote user authentication · Biometric

1 Introduction

A distributed network of autonomous sensors that can collect information related to environmental or physical conditions is called wireless sensor network(WSN). Thanks to its easiness and inexpensive deployment capabilities, WSN is applicable to numerous scientific and technological areas: Environmental monitoring, a smart grid, health care, security and surveillance, an earthquake, fire and other human activities and physical and environmental phenomena. For these reasons, a security of WSN is as important as its variety of applications. In particular, if user's personal information is contained, it should not be exposed to others.

* Corresponding author.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_5

WSN systems consist of three entities: a user interface, sensor nodes that measure physical or environmental conditions, and gateway nodes that forward information received from sensor nodes to a central server. WSN should provide simplicity and efficiency to users and must also be secure. Even if intercepting data packets sent from the WSN, an unauthorized user should not know any private information, such as the user's identity. Furthermore, any user should not be able to be authenticated as another user. However, the problem we found is that these conditions do not hold in Wu et al.'s scheme [1].

1.1 Related Work

In 2004, Watro et al. [2] suggested a user authentication scheme using the RSA and Diffie-Hellman key exchange algorithm. In 2009, the first two-factor user authentication scheme for WSNs was introduced by Das [3]. In their scheme, to pass a gateway node's checking steps, a legitimate user should have not only a password but also a smart card. This mechanism had been applied for many years in client/server networks [4–7]. However, He et al. [8] have discovered that Das' scheme was susceptible to several attacks such as an insider attack, impersonation attack, and it had lack of mutual authentication. For these reasons, they proposed the improved scheme. Unfortunately, Kumar et al. [9] mentioned that there were several vulnerabilities such as information leakage, no session key agreement, no user anonymity, and no mutual authentication in the scheme [8] In 2011, Yeh et al. [10] suggested the first two-factor user authentication scheme for WSNs using elliptic curve cryptosystem. In addition, In 2013, Xue et al. [11] proposed a temporal-credential-based authentication scheme for WSNs. In fact, a temporal credential is a result from hashing the shared key between the user and the gateway, the user's identity, and the expiration time of the temporal credential. However, it is proved by Jiang et al. [12] that the scheme [11] was insecure to the identity guessing attack, insider and tracking attacks, and off-line password guessing attack. As a result, they proposed a new mechanism in the scheme [12].

In 2014, Das [13] explained that there are some significant problems in Jiang et al.'s two-factor user authentication method [12], such as vulnerability of insider attack, lack of no formal security verification, and de-synchronization attacks, so they suggested a new three-factor user authentication scheme. In 2015, Das also introduced two three-factor authentication schemes in [14, 15], individually. In 2018, however, Wu et al. [1] found that Das' schemes [13–15] are still vulnerable. The scheme [13] was susceptible to off-line password guessing and de-synchronization attacks, and schemes [14, 15] could not withstand the off-line password guessing, user impersonation attacks. Wu et al. [1] designed an improved user authentication scheme using elliptic curve cryptography(ECC) which has been applied for WSN recently.

Unfortunately, we have found that Wu et al. [1]'s scheme is still unreliable. To be specific, Wu et al.'s scheme is exposed to the outsider, user impersonation attacks and do not satisfy user anonymity.

1.2 Organization of our paper

The rest of the paper is summarized as follows. In Section 2, we provide some preliminary knowledge such as ECC, fuzzy extractor and threat model. In addition, we review Wu et al.'s scheme of [1] in Section 3. In Section 4, we specify some vulnerabilities in Wu et al.'s scheme [1]. At last, the conclusion is shown in Section 5.

2 Preliminary Knowledge

This section describes the basic backgrounds of the elliptic curves and contents of the fuzzy extractor which are used in Wu et al.'s scheme [1] and threat model.

2.1 Elliptic Curve Cryptosystem

Elliptic curve cryptosystem(ECC) is the most frequently used password in modern passwords and has strong security characteristics. The elliptic curve cryptosystem created by Victor Miller [16] and Neal Kobiltz [17] in 1985 and 1987. It has the following form:

$$y^2 = x^3 + ax + b \quad mod\ p \qquad a, b \in F_p \tag{1}$$

Equation 1 is an equation of elliptic curve cryptosystem on the field F_p. The following conditions must be met to ensure safety.

$$4a^3 + 27b^2 \neq 0 \quad mod\ p \tag{2}$$

Equation 2 guarantees non-singular of an elliptic curve. In other words, using this elliptic curve equation 2, the following safety is guaranteed. We assume that P is the point on the elliptic curve, xP is the computation of P times x, yP is the computation of P times y, and xyP is the computation of P times xy.

1. *Elliptic Curve Decisional Diffie-Hellman Problem:* Given xP, yP it is impossible to find xyP.
2. *Elliptic Curve Computational Diffie-Hellman Problem:* Given xyP, it is impossible to find xP, yP.
3. *Elliptic Curve Discrete Logarithm Problem:* Given P, xP it is impossible to find x.

2.2 Fuzzy Extractor

User's biometric information is very important and sensitive information. In general, human biometrics can be perceived as a different result. The fuzzy extractor retrieves everybody's biometrics with a random arbitrary bit stream. User can get owns a secret string using error tolerance through the fuzzy extractor. Based on Refs [18, 19], the fuzzy extractor is worked through two processes (*Gen*, *Rep*) as follows:

$$Gen(B) \rightarrow \langle \alpha, \beta \rangle \tag{3}$$

$$Rep(B^*, \beta) = \alpha \text{ if } BIO^* \text{ is reasonably close to BIO} \tag{4}$$

From above equations, *Gen* is a probabilistic generation function using biometrics B , and extracts string $\alpha \in \{0,1\}^k$ and auxiliary string $\beta \in \{0,1\}^*$. On the other hand, *Rep* is a deterministic reproduction function that recovers α from β and any vector BIO^* that is reasonably close to BIO. For further details of the fuzzy extractor, see [20].

2.3 Threat Model

In this subsection, we describe some threat model [21] and consider constructing the assumptions of the threat model are shown as follows:

1. The attacker $\mathcal{A}$ could be either a user, sensor, or gateway. Any certified user can act as an attacker.
2. $\mathcal{A}$ could intercept or snoop all communication messages in a public channel so that $\mathcal{A}$ could steal any messages communicated between a user and sensor or gateway.
3. $\mathcal{A}$ has the capability of modifying, rerouting or deleting the intercepted message.
4. Using a side channel attack, stored parameters can be drawn from the smart card.

3 Review of Wu et al.'s scheme

In this section, we review Wu et al.'s scheme [1] to do the cryptanalysis on their scheme. The scheme consists of four phases as follows: registration phase, login phase, authentication phase, and password change phase. As schemes in [19], the scheme employs the ECC. GWN, first, produces G on $E(F_p)$ using a generator P and a large prime order n. GWN, then, chooses a private key x of which length is the security length l_s and two cryptographic hash functions $h(\cdot)$ and $h_1(\cdot)$. They are considered that the all the random generated numbers should reach the length l_s. The notations used in Wu et al.'s scheme are written in Table 1.

3.1 Registration Phase

This phase consists of two parts: user registration and sensor registration.

Table 1. Notations used in Wu et al.'s scheme.

Notations	Description
U_i	The i-th user
S_j, SID_j	The j-th sensor and its identity
ID_i	U_i's identity
PW_i	U_i's Password
B_i	U_i's biometric information
$\mathcal{A}$	The malicious attacker
x	Private key of GWN
r_i	U_i's randomly generated number
$h(\cdot)$, $h_1(\cdot)$	One-way hash function
$X\|\|Y$	Concatenation operation
$\oplus$	Bitwise XOR operation
$E(F_p)$	A collection of points on an elliptic curve over a finite field F_p
P	A point generator in F_p with a large prime order n
G	A cyclic addition group with point generator P
sk_u, sk_s	The session key generated by U_i and S_j respectively.
l_s	Security length variable

User registration

1. An user U_i, first, decides his/her identity ID_i and password PW_i with a randomly generated number r_i, imprints B_i over a device for biometrics collection, and calculates $Gen(B_i) = (R_i, P_{bi})$, $DID_i = h(ID_i \parallel r_i)$ and $HPW_i = h(PW_i \parallel r_i \parallel R_i)$. He/she, then, transmits the registration request $\{ID_i, DID_i\}$ to the gateway node GWN in the secure channel.
2. After obtaining the registration request from the U_i, GWN computes $B_1' = h(DID_i \parallel x)$ where the value x is a secret key of GWN, produces a smart card for U_i holding $h(\cdot)$, $h_1(\cdot)$, P, and stores ID_i in its database. GWN then delivers the smart card with B_1' to the U_i secretly.
3. After taking the smart card with B_1' from the GWN, U_i computes $B_1 = B_1' \oplus HPW_i$ and $B_2 = h(ID_i \parallel R_i \parallel PW_i) \oplus r_i$ with storing B_1, B_2, P and P_{bi} into the smart card.

Sensor registration

1. GWN picks an identity $SIDj$ for each new sensor node S_j, calculates $c_j = h(SID_j \parallel x)$, and sends $\{SID_j, c_j\}$ to S_j.
2. S_j stores P, SID_j and c_j, and follows the WSN.

3.2 Login Phase

1. U_i enters ID_i, PW_i and B_i'. The smart card generates $Rep(B_i', P_{bi}) = R_i$, $r_i = B_2 \oplus h(ID_i \parallel R_i \parallel PW_i)$, $HPW_i = h(PW_i \parallel r_i \parallel R_i)$ and $DID_i = h(ID_i \parallel r_i)$.

2. The smart card produces randomly generated numbers r_i^{new}, e_i and $\alpha \in [1, n-1]$, and chooses a special sensor SID_j. The smart card then computes $DID_i^{new} = h(ID_i \parallel r_i^{new})$, $C_1 = B_1 \oplus HPW_i \oplus e_i$, $C_2 = \alpha P$, $C_3 = h(e_i) \oplus DID_i^{new}$, $Z_i = ID_i \oplus h(e_i \parallel DID_i)$ and $C_4 = h(ID_i \parallel e_i \parallel DID_i \parallel DID_i^{new} \parallel C_2 \parallel SID_j)$. The value C_4 is used for checking the identities' integrity and the user side's new data and verifying the source of the message M_1.
3. U_i sends the login request messages $M_1 = \{C_1, C_2, C_3, C_4, Z_i, DID_i, SID_j\}$ to GWN.

3.3 Authentication Phase

1. After accepting the login request messages M_1 from the user U_i, GWN first computes $e_i = C_1 \oplus h(DID_i \parallel x)$, $DID_i^{new} = C_3 \oplus h(e_i)$ and $ID_i = Z_i \oplus h(e_i \parallel DID_i)$, and checks the validity of ID_i and $C_4 \overset{?}{=} h(ID_i \parallel e_i \parallel DID_i \parallel DID_i^{new} \parallel C_2 \parallel SID_j)$. If either fails, GWN terminates the session. If authentication attempts fail three times in a row in a defined time span, GWN will freeze the U_i's account; otherwise, GWN calculates $c_j = h(SID_j \parallel x)$ and $C_5 = h(c_j \parallel DID_j \parallel SID_j \parallel C_2)$ and sends $M_2 = \{C_2, C_5, DID_i\}$ to the sensor node S_j. The value C_5 is used for checking the integrity of the strings including c_j and the data that can make the sensor S_j to obtain the correct data for computing the session key. In addition, C_5 is used for verifying the source of M_2.
2. S_j checks $C_5 \overset{?}{=} h(c_j \parallel DID_i \parallel SID_j \parallel C_2)$ with its identity SID_j. If this does not hold, S_j will disconnect the session. S_j, then, selects $\beta \in [1, n-1]$, and computes $C_6 = \beta P$, $sk_s = \beta C_2$, $C_7 = h_1(C_2 \parallel C_6 \parallel sk_s \parallel DID_i \parallel SID_j)$ and $C_8 = h(DID_i \parallel SID_j \parallel c_j)$. The major role of C_7 is to check the session key's integrity and C_6's integrity, which is the part used by U_i to compute the session key. Furthermore, both C_7 and C_8 are used to verifying the source of M_3. In the end, S_j transmits $M_3 = \{C_6, C_7, C_8\}$ to GWN.
3. GWN checks $C_8 \overset{?}{=} h(DID_i \parallel SID_j \parallel c_j)$. If this does not satisfy, GWN disconnect the session; otherwise, GWN computes $C_9 = h(DID_i^{new} \parallel x) \oplus h(DID_i \parallel e_i)$ and $C_{10} = h(ID_i \parallel SID_j \parallel DID_i \parallel DID_i^{new} \parallel e_i \parallel C_9)$. The value C_{10} is to verify the source of the message M_4. Finally, GWN sends the message $M_4 = \{C_6, C_7, C_9, C_{10}\}$ to U_i.
4. U_i checks $C_{10} \overset{?}{=} h(ID_i \parallel SID_j \parallel DID_i \parallel DID_i^{new} \parallel e_i \parallel C_9)$. U_i then computes the session key $sk_u = \alpha C_6$, and checks $C_7 \overset{?}{=} h_1(C_2 \parallel C_6 \parallel sk_u \parallel DID_i \parallel SID_j)$. If this does not satisfy, U_i terminates the session. After that, U_i calculate $HPW_i^{new} = h(PW_i \parallel r_i^{new} \parallel R_i)$, $B_1^{new} = C_9 \oplus h(DID_i \parallel e_i) \oplus HPW_i^{new}$ and $B_2^{new} = h(ID_i \parallel R_i \parallel PW_i) \oplus r_i^{new}$, and replaces ($B_1$, B_2) with (B_1^{new}, B_2^{new}) in the smart card individually.

3.4 Password and Biometrics Change Phase

1. This step is same as the first step of Login phase.

2. The smart card produces random generated numbers r_i^{new} and e_i, calculates DID_i^{new}, C_1, C_3, Z_i and $C_{11} = h(ID_i \parallel e_i \parallel DID_i \parallel DID_i^{new})$, and sends $M_5 = \{C_1, C_3, Z_i, C_{11}, DID_i\}$ with a password change request to GWN. The value C_{11} is similar to C_4 and it is used for checking the integrity of the identities and verifying the source of M_5.
3. GWN acquires e_i, ID_i and DID_i^{new} as first step of the authentication phase, and determines ID_i and $C_{11} \overset{?}{=} h(ID_i \parallel e_i \parallel DID_i \parallel DID_i^{new})$. If this does not satisfy, GWN disconnects the session; otherwise, GWN generates $C_9 = h(DID_i^{new} \parallel x) \oplus h(DID_i \parallel e_i)$ and $C_{12} = h(ID_i \parallel DID_i \parallel DID_i^{new} \parallel e_i \parallel C_9)$ and sends $M_6 = \{C_9, C_{12}\}$ and a grant to U_i. Here C_{12} is to verify the source of M_6.
4. U_i checks $C_{12} \overset{?}{=} h(ID_i \parallel DID_i \parallel DID_i^{new} \parallel e_i \parallel C_9)$. If it is incorrect, U_i disconnects this session; otherwise, U_i inputs a new password PW_i^{new} and a new biometric information B_i^{new}. The smart card then computes $Gen(B_i^{new}) = (R_i^{new}, P_{bi}^{new})$, $HPW_i^{new2} = h(PW_i^{new} \parallel r_i^{new} \parallel R_i^{new})$, $B1^{new2} = C_9 \oplus h(DID_i \parallel e_i) \oplus HPW_i^{new2}$ and $B_2^{new2} = h(ID_i \parallel R_i^{new} \parallel PW_i^{new}) \oplus r_i^{new}$. Finally, U_i substitutes $(B_1^{new2}, B_2^{new2}, P_{bi}^{new2})$ for (B_1, B_2, P_{bi}) in the smart card individually.

4 Security Weaknesses of Wu et al.'s scheme

In this section, we prove that Wu et al.'s scheme [1] has some security exposure. The following issues have been found and their specific descriptions are given below.

4.1 Outsider Attack

1. An attacker $\mathcal{A}$ who is the legitimate user and owns a his/her own smart card can extract the $\{B_{1\mathcal{A}}, B_{2\mathcal{A}}, P, P_{b\mathcal{A}}\}$ from his/her smart card.
2. $\mathcal{A}$ can thus get $h(DID_{\mathcal{A}} \parallel x) = B_{1\mathcal{A}} \oplus HPW_{\mathcal{A}}$, and use this value for other attacks. Because, this value is an important value that identifies the user on the gateway node side. $h(DID_A \parallel x)$ will be used in Section 4.2 and Section 4.3.

4.2 User Impersonation Attack

An attacker $\mathcal{A}$ can pretend any user using his/her information and other user's identity alone. We assume that the victim is user U_i at this time. The specific method is shown as follows in detailed.

1. The attacker $\mathcal{A}$ selects any identity ID_i.
2. $\mathcal{A}$ generates random numbers $r_{\mathcal{A}}^{new}$, $e_{\mathcal{A}}$, and $\alpha_{\mathcal{A}} \in [1, n-1]$, and chooses a special sensor SID_j. $\mathcal{A}$ then computes $DID_{\mathcal{A}}^{new} = h(ID_{\mathcal{A}} \parallel r_{\mathcal{A}}^{new})$, $C_{1\mathcal{A}} = B_{1\mathcal{A}} \oplus HPW_{\mathcal{A}} \oplus e_{\mathcal{A}}$, $C_{2\mathcal{A}} = \alpha_{\mathcal{A}} P$, $C_{3\mathcal{A}} = h(e_{\mathcal{A}}) \oplus DID_{\mathcal{A}}^{new}$, $Z_{\mathcal{A}} = ID_i \oplus h(e_{\mathcal{A}} \parallel DID_{\mathcal{A}})$ and $C_{4\mathcal{A}} = h(ID_i \parallel e_{\mathcal{A}} \parallel DID_{\mathcal{A}} \parallel DID_{\mathcal{A}}^{new} \parallel C_{2\mathcal{A}} \parallel SID_j)$. $C_{4\mathcal{A}}$ is used for checking the integrity of the identities and the new data produced on the user side and verifying the source of $M_{1\mathcal{A}}$.

3. A transmits the login request $M_{1\mathcal{A}} = \{C_{1\mathcal{A}}, C_{2\mathcal{A}}, C_{3\mathcal{A}}, C_{4\mathcal{A}}, Z_{\mathcal{A}}, DID_{\mathcal{A}}, SID_j\}$ to the gateway node GWN.
4. After obtaining the login request from the $\mathcal{A}$, GWN, first, calculates $e_{\mathcal{A}} = C_{1\mathcal{A}} \oplus h(DID_{\mathcal{A}} \parallel x)$, $DID_{\mathcal{A}}^{new} = C_{3\mathcal{A}} \oplus h(e_{\mathcal{A}})$ and $ID_i = Z_{\mathcal{A}} \oplus h(e_{\mathcal{A}} \parallel DID_{\mathcal{A}})$, and checks the validity of ID_i and $C_{4\mathcal{A}} \overset{?}{=} h(ID_i \parallel e_{\mathcal{A}} \parallel DID_{\mathcal{A}} \parallel DID_{\mathcal{A}}^{new} \parallel C_{2\mathcal{A}} \parallel SID_j)$. GWN proceeds the scheme without any detection. Unfortunately, the GWN misunderstand that he/she is communicating with the valid victim U_i.

As a result, the attacker A will be verified as user U_i by user GWN. Therefore, the user impersonation attack is succeed.

4.3 No User Anonymity

The attacker $\mathcal{A}$ can extract the identity of U_i from the login request message M_i of U_i. Assume that $\mathcal{A}$ eavesdrops the login request message $M_1 = \{C_1, C_2, C_3, C_4, Z_i, DID_i, SID_j\}$ of U_i. The details are as follows.

1. The attacker $\mathcal{A}$ first generates randomly generated numbers $r_{\mathcal{A}}^{new}$, $e_{\mathcal{A}}$, and $\alpha_{\mathcal{A}} \in [1, n-1]$, and chooses a special sensor SID_j. $C_{1\mathcal{A}} = B_{1\mathcal{A}} \oplus HPW_{\mathcal{A}} \oplus e_{\mathcal{A}}$, $C_{2\mathcal{A}} = \alpha_{\mathcal{A}} P$, $C_{3\mathcal{A}} = h(e_{\mathcal{A}}) \oplus DID_i$, $Z_{\mathcal{A}} = ID_{\mathcal{A}} \oplus h(e_{\mathcal{A}} \parallel DID_{\mathcal{A}})$ and $C_{4\mathcal{A}} = h(ID_{\mathcal{A}} \parallel e_{\mathcal{A}} \parallel DID_{\mathcal{A}} \parallel DID_i \parallel C_{2\mathcal{A}} \parallel SID_j)$.
2. A sends the login request message $M_{1\mathcal{A}} = \{C_{1\mathcal{A}}, C_{2\mathcal{A}}, C_{3\mathcal{A}}, C_{4\mathcal{A}}, Z_{\mathcal{A}}, DID_{\mathcal{A}}, SID_j\}$ to the gateway node GWN.
3. After getting the login request message from the $\mathcal{A}$, GWN calculates $e_{\mathcal{A}} = C_{1\mathcal{A}} \oplus h(DID_{\mathcal{A}} \parallel x)$, $DID_i = C_{3\mathcal{A}} \oplus h(e_{\mathcal{A}})$ and $ID_{\mathcal{A}} = Z_{\mathcal{A}} \oplus h(e_{\mathcal{A}} \parallel DID_{\mathcal{A}})$, and checks the validity of $ID_{\mathcal{A}}$ and $C_{4\mathcal{A}} \overset{?}{=} h(ID_{\mathcal{A}} \parallel e_{\mathcal{A}} \parallel DID_{\mathcal{A}} \parallel DID_i \parallel C_{2\mathcal{A}} \parallel SID_j)$. GWN then computes $c_j = h(SID_j \parallel x)$ and $C_{5\mathcal{A}} = h(c_j \parallel DID_j \parallel SID_j \parallel C_{2\mathcal{A}})$ and sends $M_{2\mathcal{A}} = \{C_{2\mathcal{A}}, C_{5\mathcal{A}}, DID_{\mathcal{A}}\}$ to the sensor node S_j.
4. S_j checks $C_{5\mathcal{A}} \overset{?}{=} h(c_j \parallel DID_{\mathcal{A}} \parallel SID_j \parallel C_{2\mathcal{A}})$ with its identity SID_j. If it is incorrect, S_j terminates the session. S_j then selects $\beta_{\mathcal{A}} \in [1,\ n-1]$ and computes $C_{6\mathcal{A}} = \beta_{\mathcal{A}} P$, $sk_s = \beta_{\mathcal{A}} C_{2\mathcal{A}}$, $C_{7\mathcal{A}} = h_1(C_{2\mathcal{A}} \parallel C_{6\mathcal{A}} \parallel sk_s \parallel DID_{\mathcal{A}} \parallel SID_j)$ and $C_{8\mathcal{A}} = h(DID_{\mathcal{A}} \parallel SID_j \parallel c_j)$. S_j sends $M_{3\mathcal{A}} = \{C_{6\mathcal{A}}, C_{7\mathcal{A}}, C_{8\mathcal{A}}\}$ to GWN.
5. GWN checks $C_{8\mathcal{A}} \overset{?}{=} h(DID_{\mathcal{A}} \parallel SID_j \parallel c_j)$. If this does not hold, GWN terminates the session; otherwise, GWN calculates $C_{9\mathcal{A}} = h(DID_i \parallel x) \oplus h(DID_{\mathcal{A}} \parallel e_{\mathcal{A}})$ and $C_{10\mathcal{A}} = h(ID_{\mathcal{A}} \parallel SID_j \parallel DID_{\mathcal{A}} \parallel DID_i \parallel e_{\mathcal{A}} \parallel C_{9\mathcal{A}})$. Finally GWN sends the message $M_{4\mathcal{A}} = \{C_{6\mathcal{A}}, C_{7\mathcal{A}}, C_{9\mathcal{A}}, C_{10\mathcal{A}}\}$ to attacker $\mathcal{A}$.
6. $\mathcal{A}$ computes $h(DID_i \parallel x) = h(DID_{\mathcal{A}} \parallel e_{\mathcal{A}}) \oplus C_{9\mathcal{A}}$. Now $\mathcal{A}$ can compute $e_i = C_1 \oplus h(DID_i \parallel x)$. Finally, $\mathcal{A}$ can find $ID_i = h(e_i \parallel DID_i) \oplus Z_i$.

As a result, this result shows that Wu et al.'s scheme does not satisfy user anonymity.

5 Conclusions

In this paper, we reviewed Wu et al.'s three-factor user authentication scheme for WSN and demonstrated that outsider attack is still possible in Wu et al.'s scheme. The outsider attack could be used to pull out security-critical information. As a result, It brings about exposure of session key, user impersonation attack and no user anonymity. For these reasons, it is not secure to use their authentication scheme. Especially, ID must not exposed as an XOR to prevent user impersonation attack. Future research will need to be done in a way that will complement it. Finally, our further research would be focused on proposing an advanced user authentication scheme which can handle with these problems.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2010-0020210)

References

1. Wu, F., Xu, L., Kumari, S., Li, X.: An Improved and Provably Secure Three-Factor User Authentication Scheme for Wireless Sensor Networks. Peer-to-Peer Networking and Applications **11**(1), 1–20 (2018)
2. Watro, R., Kong, D., Cuti, Sf., Gardiner, C., Lynn, C., Kruus, P.: Tinypk: Securing Sensor Networks with Public Key Technology. In: Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks. ACM, 59–64 (2004)
3. Das, M.: Two-Factor User Authentication in Wireless Sensor Networks. IEEE transactions on wireless communications **8**(3), 1086-1090 (2009)
4. Choi, Y., Lee, Y., Won, D.: Security Improvement on Biometric Based Authentication Scheme for Wireless Sensor Networks Using Fuzzy Extraction. International Journal of Distributed Sensor Networks **2016**, 1-16 (2016)
5. Kim, J., Moon, J., Jung, J., Won, D.: Security Analysis and Improvements of Session Key Establishment for Clustered Sensor Networks. Journal of Sensors **2016**, 1-17 (2016)
6. Kang, D., Jung, J., Mun, J., Lee, D., Choi, Y., Won, D.: Efficient and Robust User Authentication Scheme that Achieve User Anonymity with a Markov Chain. Security and Communication Networks **9**(11), 1462-1476 (2016)
7. Jung, J., Kim, J., Choi, Y., Won, D.: An Anonymous User Authentication and Key Agreement Scheme Based on a Symmetric Cryptosystem in Wireless Sensor Networks. Sensors **16**(8), 1-30 (2016)
8. He, D., Gao, Y., Chan, S., Chen, C., Bu, J.: An enhanced two-factor user authentication scheme in wireless sensor networks. Ad hoc and Sensor wireless network **10**(4), 361-371 (2010)
9. Kumar, P., Lee, H. J.: Cryptanalysis on two user authentication protocols using smart card for wireless sensor networks. IEEE Wireless advanced (WiAd), 241-245 (2011)

10. Yeh, H. L., Chen, T. H., Liu, P. C., Kim, T. H., Wei, H. W.: A secured authentication protocol for wireless sensor networks using elliptic curves cryptography. Sensors, **11**(5), 4767-4779 (2011)
11. Xue, K., Ma, C., Hong, P., Ding, R.: A temporal-credential-based mutual authentication and key agreement scheme for wireless sensor networks. Journal of Network and Computer Applications, **36**(1), 316-323 (2013)
12. Jiang, Q., Ma, J., Lu, X., Tian, Y.: An efficient two-factor user authentication scheme with unlinkability for wireless sensor networks. Peer-to-peer Networking and Applications, **8**(6), 1070-1081 (2015)
13. Das, A. K.: A secure and robust temporal credential-based three-factor user authentication scheme for wireless sensor networks. Peer-to-peer Networking and Applications, **9**(1), 223-244 (2016)
14. Das, A. K.: A secure and effective biometricbased user authentication scheme for wireless sensor networks using smart card and fuzzy extractor. International Journal of Communication Systems, **30**(1) (2017)
15. Das, A. K.: A secure and efficient user anonymity-preserving three-factor authentication protocol for large-scale distributed wireless sensor networks. Wireless Personal Communications, **82**(3), 1377-1404 (2015)
16. Miller, V.: Uses of Elliptic Curves in Cryptography. In: Advances in Cryptology Crypto **218**, 417–426 (1986)
17. Koblitz, N.: Elliptic curve cryptosystems. Mathematics of computation **48**, 203–209 (1987)
18. Dodis, Y., Kanukurthi, B., Katz, J., Smith, A.: Robust fuzzy extractors and authenticated key agreement from close secrets. IEEE Transactions on Information Theory **58**, 6207-6222 (2013)
19. Das, A.: A Secure and Effective Biometric-based User Authentication Scheme for Wireless Sensor Networks using Smart Card and Fuzzy Extractor. International Journal of Communication Systems **2015**, 1-25 (2015)
20. Dodis, Y., Reyzin, L., Smith, A.: Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. In: Proceedings on the International Conference on the Theory and Applications of Cryptographic Techniques, 523–540 (2004)
21. Moon, J., Choi, Y., Jung, J., Won, D.: An Improvement of Robust Biometrics-based Authentication and Key Agreement Scheme for Multi-Server Environments using Smart Cards. PLoS One **10**, 1–15 (2015)

User Profiling in Anomaly Detection of Authorization Logs

Zahedeh Zamanian[1], Ali Feizollah[1] , Nor Badrul Anuar[1], Miss Laiha Binti Mat Kiah[1], Karanam Srikanth[2], Sudhindra Kumar[2]

[1] Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia
[2] NextLabs (Malaysia) Sdn Bhd, No.308-1st Floor, Jalan S2 B13,
Seksyen B, Uptown Avenue Seremban 2,
70300 Seremban, Negeri Sembilan, Malaysia

ali.feizollah@siswa.um.edu.my

Abstract. In digital age, the valuable asset of every company is their data. They contain personal information, companies and industries data, sensitive government communications and a lot of more. With the rapid development in IT technology, accessing the network become cheaper and easier. As a result, organizations are more vulnerable to both insiders and outsider threat. This work proposes user profiling in anomaly detection and analysis of log authorization. This method enables companies to assess each user's activities and detect slight deviation from their usual pattern. To evaluate this method, we obtained a private dataset from NextLabs Company, and the CERT dataset that is a public dataset. We used random forest for this system and presented the results. The result shows that the algorithm achieved 97.81% of accuracy.

Keywords: User Profiling, Anomaly Detection, Insider Intruder.

1 Introduction

Whenever there are valuable assets, there will be thieves who want to steal those assets. In the past this concept applied more to valuable physical objects such as gold, jewelry and money, but in the last decades with dawn of digital, this concept covers broader areas. Digital valuable objects could be personal information, companies and industries data, sensitive government communications and a lot of more. With the rapid development in IT technology, accessing the network become cheaper and easier. As a result, organizations are more vulnerable to both insider and outsider threat. Thus, data protection and securing networks become vitally important [1]. A report by the FBI's Internet Crime Complaint Center (IC3) in 2016 shows that during 2012 until 2016 approximately 280,000 internet scams complaints per year was received by IC3. Fig 1 illustrates over that time period, IC3 received a total of 1,408,849 complaints, and a total reported loss of USD 4.63 billion. The complaints address a wide range of Internet scams affecting victims across the globe [2].

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_6

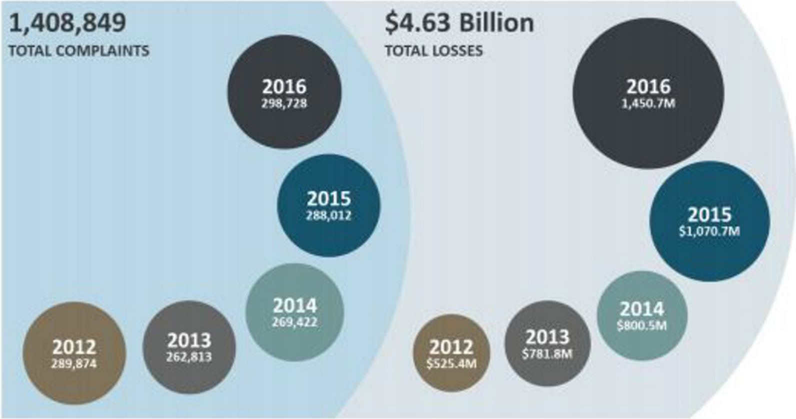


Fig.1. Number of complain and money loss from 2012 to 2016 based on IC3 report

Moreover, based on ISACA and RSA Conference Survey 2016, 42% and 64% of respondents agreed that rapid advancement of Artificial Intelligent (AI) will lead to increase of the cybersecurity/information security risk in short term and long term, respectively. The report data reveal that 20% of companies are dealing with insider damage and theft of intellectual property at least quarterly [3]. Therefore, it is crucial for companies to realize threats which influence their assets and the areas, which each threat could affect [4]. Intrusion detection (ID) method could be used to detect the threats and alert the security accordingly. Intrusion detection concept was introduced by Anderson in 1980. He proposed ID as a method to identify when a system was compromised [5]. To detect intrusions, analyzing of log files could be used to identify behaviors that indicate misuse of the system [6].

The current literature have focused on intrusion detection based on analyzing all data [7] [8]. In this method, data of all users in a company are gathered and analysis is performed. Then, a model is built based on all users, that is used for anomaly detection. In this work, we propose studying and profiling users' behavior. Based on this method, models are built for users that represent their behavior in interaction with the system. A slight deviation in such behavior raises suspicion, whereas in previous detection system behavior of all users, their behaviors are analyzed together.

The rest of the paper is presented as following. Section 2 discusses related works to this study. It includes their method and results along with their limitations. Section 3 details the method and experimental setup. It also explains sources of the included datasets and their description. Section 4 presents results and discusses them to evaluate the proposed method. Section 5 concludes this work by pointing out bold points.

2 Related Works

Anomaly detection has been an important research problem in security analysis, therefore development of methods that can detect malicious insider behavior with high accuracy and low false alarm is vital [9]. In this problem layout, McGough *et al* [7] designed a system to identify anomalous behavior of user by comparing of individual user's activities against their own routine profile, as well as against the organization's rule. They applied two independent approaches of machine learning and Statistical Analyzer on data. Then results from these two parts combined together to form consensus which then mapped to a risk score. Their system showed high accuracy, low false positive and minimum effect on the existing computing and network resources in terms of memory and CPU usage.

Bhattacharjee et al proposed a graph-based method that can investigate user behavior from two perspectives: (a) anomaly with reference to the normal activities of individual user which has been observed in a prolonged period of time, and (b) finding the relationship between user and his colleagues with similar roles/profiles. They utilized CMU-CERT dataset in unsupervised manner. In their model, Boykov Kolmogorov algorithm was used and the result compared with different algorithms including Single Model One-Class SVM, Individual Profile Analysis, k-User Clustering and Maximum Clique (MC). Their proposed model evaluated by evaluation metrics Area-Under-Curve (AUC) that showed impressive improvement compare to other algorithms [8]. Log data are considered as high-dimensional data which contain irrelevant and redundant features. Feature selection methods can be applied to reduce dimensionality, decrease training time and enhance learning performance [10].

In [11] Legg et al. offered an automated system that construct tree structured profiles based on individual user activity and combined role activity. This method helped them to attain consistent features which provide description of the user's behavior. They reduced high dimensionality of this feature set by using principal component analysis (PCA) and compute anomaly scores based on Mahalanobis distance anomaly metrics. Their system was tested on synthetic dataset which ten malicious data injected. Their system performed well for identifying these attacks.

In a similar line, Agrafiotis et al [12] applied same model as offered by Legg et al but they used real-world data set from multinational organization. Moreover, their approach abided the ethical and privacy concerns. Their result showed high accuracy and low false alarm. Although finding a sequence is a common choice for modeling activities and events through time but catching anomalous sequence in a dataset is not an easy task. One of the algorithms that has ability to recognize temporal pattern and widely has been used is Hidden Markov Models (HMM).

Rashid et al [13] proposed a model based on HMM to identify insider threat in CERT dataset. They tried to model user's normal behavior as a week-long sequence. Their modeled showed accurate result with low false alarm. Although author mentioned using shorter time frame for instance a day long sequences could build a more accurate model of employee's daily behavior. Moreover, their system was trained based on first 5 weeks and thus it is not able to detect insider threats amongst short-term users such as contractors whose are the real threats. The user profiling was not mentioned in the above works, despite being an effective approach.

3 The Method and Experimental Setup

This method takes data for each user such as daily activities in a month. This data are gathered and a model is built for that user to create a pattern of the user's activity. Once the user is involved in an activity, it is compared to his/her behavioral pattern. In case it is deviated from the normal behavior, it is flagged suspicious. Since a slight deviation could be a false alarm, this activity is compare to behavioral model of all system's users. This way, the false alarm is kept at the minimum. Fig 2 shows various activities of a user and possibility of using those features in anomaly detection and detection of insider threat.

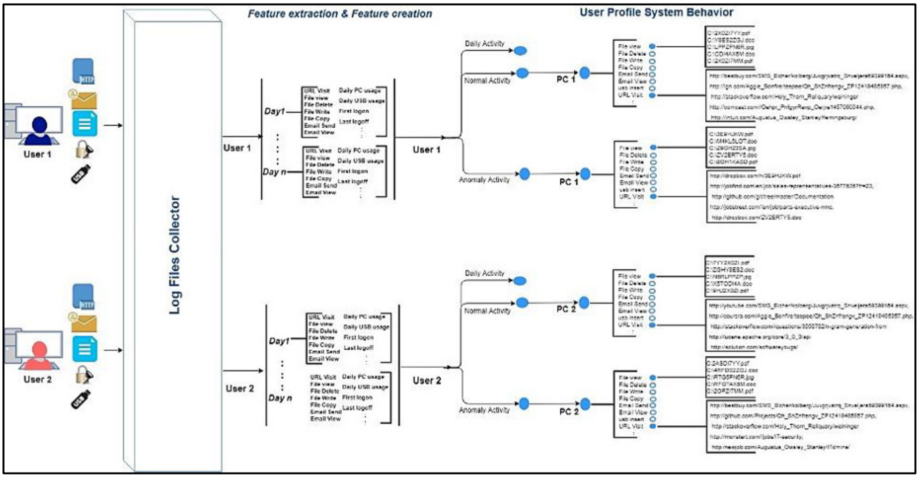


Fig.2. User Profiling Method in Authorization Logs

In order to test this method, we obtained a private and a public dataset. A set of data was acquired from NextLabs, which is a leading security company in Dynamic Authorization technology with Attribute Based Access Control (ABAC). This dataset is a collection of access logs of different users performing various activities, such as time and date, file accessed, resource used, etc.

A public dataset is also used for this experiment. The CERT was acquired, which is a collection of users' activity based on daily basis. Their use of files and resources are logged and time and date of system usage.

This work uses random forest algorithm to train, test, and generate a model for anomaly detection in log files. Random forest is an ensemble learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally cheap process. If we can build many small, weak decision trees in parallel, we can then combine the trees to form a single, strong learner by averaging or taking the majority vote. In practice, random forests are often found to be the most accurate learning algorithms to date. The random forest algorithm uses the bagging technique for building an ensemble of decision trees. Bagging is known to reduce the variance of the algorithm [14, 15].

4 Results and Discussion

This section presents results of the experiment. The data were divided into 70% training and 30% test data. The training data are used to train the algorithm. Then, the learned model is tested using test data. The test data is fed to the model as new data to measure how the algorithm is trained. The results are measured in terms of accuracy, which is number of correctly classified data over all of data.

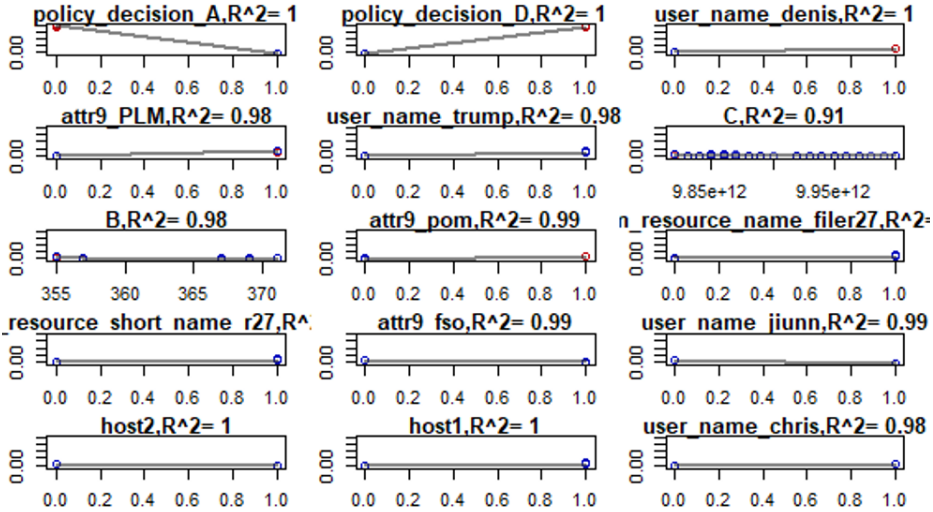


Fig.3. Variable Importance in Random Forest

Fig 3 shows importance of features. The figure shows different features in the database. Resource name is name of resource that a user utilizes. Host is name of individual machine that a user works on. This way, we know what features are important for a user when building the model. Policy decision has higher rank compare to other features, since it gives final decision whether a user is allowed or denied access.

After identifying important features, they are fed to the random forest algorithm. Table 1 shows results of the experiment.

Table 1. Experiment Result

Accuracy	Error
97.81%	2.19%

The random forest algorithm achieved 97.81% accuracy, and 2.19% of error. The error is wrongly classified data as normal or anomaly. Among various features in the dataset, policy decision is more important than others. It shows whether a user is allowed access to a resource. As illustrated in Fig 3, the random forest recognizes such importance. It is also possible to see progress of the algorithm as it trains. Fig 4 shows progress of random forest in terms of number of trees and associated error.

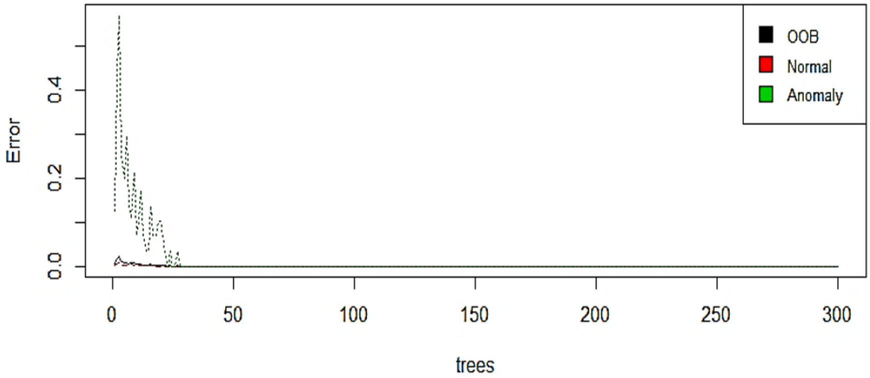


Fig.4. Random Forest Progress Graph

Fig 4 shows that the algorithm starts with high rate of error and gradually the error decreases as it learns the data. Eventually, it reaches to almost zero error rate.

5 Conclusion

Security in large companies has become a crucial issue. Insider threat is an important security issue for companies. This work proposed a method for user profiling in anomaly detection of authorization log. We used random forest algorithm for detection purpose. The private dataset was provided from NextLabs Corporation, and the public dataset was CERT. The result shows that the algorithm achieved 97.81% of accuracy.

Acknowledgement

The work described in this paper was supported by the Collaborative Agreement with NextLabs (Malaysia) Sdn Bhd (Project title: Anomaly detection in Policy Authorization Activity Logs).

References

[1] R. Prasad, "Insider Threat to Organizations in the Digital Era and Combat Strategies," presented at the Indo-US conference and workshop on "Cyber Security, Cyber Crime and Cyber Forensics, Kochi, India, 2009.

[2] S. S. Smith, "INTERNET CRIME REPORT " "FBI's Internet Crime Complaint Center "2016.

[3] C. Nexus, "State of Cybersecurity:Implications for 2016," *"An ISACA and RSA Conference Survey",* 2016.

[4] S. Bauer and E. W. N. Bernroider, "From Information Security Awareness to Reasoned Compliant Action: Analyzing Information Security Policy Compliance in a Large Banking Organization," *SIGMIS Database,* vol. 48, pp. 44-68, 2017.

[5] J. P. Anderson, "Computer security threat monitoring and surveillance," *Technical Report, James P. Anderson Company,* 1980.

[6] R. Vaarandi, M. Kont, and M. Pihelgas, "Event log analysis with the LogCluster tool," *Proceedings of Military Communications Conference MILCOM 2016-2016 IEEE,* pp. 982-987, 2016.

[7] A. S. McGough, D. Wall, J. Brennan, G. Theodoropoulos, E. Ruck-Keene, B. Arief*, et al.*, "Insider Threats: Identifying Anomalous Human Behaviour in Heterogeneous Systems Using Beneficial Intelligent Software (Ben-ware)," presented at the Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats, Denver, Colorado, USA, 2015.

[8] S. D. Bhattacharjee, J. Yuan, Z. Jiaqi, and Y.-P. Tan, "Context-aware graph-based analysis for detecting anomalous activities," presented at the Multimedia and Expo (ICME), 2017 IEEE International Conference on, 2017.

[9] K. W. Kongsg, #229, rd, N. A. Nordbotten, F. Mancini, and P. E. Engelstad, "An Internal/Insider Threat Score for Data Loss Prevention and Detection," presented at the Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics, Scottsdale, Arizona, USA, 2017.

[10] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A Survey on semi-supervised feature selection methods," *Pattern Recognition,* vol. 64, pp. 141-158, 2017/04/01/ 2017.

[11] P. A. Legg, O. Buckley, M. Goldsmith, and S. Creese, "Automated insider threat detection system using user and role-based profile assessment," *IEEE Systems Journal,* vol. 11, pp. 503-512, 2015.

[12] I. Agrafiotis, A. Erola, J. Happa, M. Goldsmith, and S. Creese, "Validating an Insider Threat Detection System: A Real Scenario Perspective," presented at the 2016 IEEE Security and Privacy Workshops (SPW), 2016.

[13] T. Rashid, I. Agrafiotis, and J. R. C. Nurse, "A New Take on Detecting Insider Threats: Exploring the Use of Hidden Markov Models," presented at the Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats, Vienna, Austria, 2016.

[14] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[15] H. Tin Kam, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 20, pp. 832-844, 1998.

Agent based integer programming framework for solving real-life curriculum-based university course timetabling

Mansour Hassani Abdalla, Joe Henry Obit, Rayner Alfred, and Jetol Bolongkikit

Knowledge Technology Research Unit, Universiti Malaysia Sabah, 88400 Kota Kinabalu,
Sabah, Malaysia
mansourabdalla22@gmail.com, joehenry@ums.edu.my,
ralfred@umse.du.my, jetol@ums.edu.my

Abstract. This research proposes an agent-based framework for solving real-life curriculum-based University Course Timetabling problems (CB-UCT) at the Universiti Malaysia Sabah, Labuan International Campus (UMSLIC). Similar to other timetabling problems, CB-UCT in UMSLIC has its own distinctive constraints and features. The proposed framework deal with the problem using a distributed Multi-Agent System (MAS) environment in which a central agent coordinates various IP agents that cooperate by sharing the best part of the solution and direct the IP agents towards more promising search space and hence improve a common global list of the solutions. All agents are incorporated with Integer programming (IP) search methodology, which is used to generate initial solution in this, regards as well. We discuss how sequential IP search methodology can be incorporated into the proposed multi-agent approach in order to conduct parallel search for CB-UCT. The agent-based IP is tested over two real-life datasets, semester 1 session 2016/2017 and semester 2 session 2016/2017. The experimental results show that the agent-based IP is able to improve the solution generated by the sequential counterpart for UMSLIC's problem instance used in the current study impressively by 12.73% and 17.89% when three and six IP agents are used respectively. Moreover, the experiment also shows that increasing the number of IP agents lead to the better results.

Keywords: Integer Programming, Multi-Agent System, Asynchronous Cooperative Search.

1 Introduction

Curriculum based university course timetabling is an interesting topic to study because neither modeling nor solving them is a straightforward task to do. This is because each problem has its own unique characteristics and variations which differ from one university to another [3]. Besides, duplicating the previous timetable does not really solve the problems as university are growing with a great pace and the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_7

teaching program is evolving towards more modular and distributed nature where students are able to choose course from other programs or even from other faculty. These fluctuations of teaching guidelines lead to the problem of constructing different timetables in every semester. Also as [3] highlight that, "poor quality course timetabling can lead to massive costs for the peoples affected by the timetable, for example students might not be able to attend all of their lessons if clashes exist". The cost here is the student need to take the course in the future semester which ultimately will result for student to extend study time. A high quality timetable is the one in which all the peoples (students, lecturers and academic departments) affected by the timetable are satisfied. However, constructing timetable which satisfies students, lecturers and academic departments is not an easy task.

Basically, all CB-UCT is associated with constraints which are different from other universities. In addition, these constraints vary from time to time and are classified into two categories which are hard constraints and soft constraints [2]. Hard constraints must be satisfied at all circumstances while soft constructs are used to determine the quality of timetable and the more the soft constraints are satisfied the better the timetable produced. In order to solve this problem for the certain real-life case of CB-UCT, we have adopted the agent-based incorporated with IP search methodology.

Generally, in past ten to twenty years ago, agent-based technology has entered the scene of software industry and proves its suitability [4]. MAS fall into the area of distributed systems, where number of entities work together to cooperatively solve given problems. [1] Pointed out that, "MAS are concerned with coordinating behavior among a pool of autonomous intelligent agents (e.g. software agents) that work in an environment". In this regard MAS is effective because it facilitates the agent to share the best part of the solution and hence guide the search process to more promising region. These agents can be cooperating to achieve common goals however generally on other systems agents are competing with each other to fulfil the delegated objectives [17] are commonly intended as computational systems where several autonomous entities called agents, interact or work together to perform some tasks[2]. Likewise, in CB-UCT, the MAS could find a high-quality and acceptable solution with minimal message passing as well. In this work, we are proposing agent-based framework incorporating IP search methodology where agents are working together by sharing the best part of the solution to achieve the delegated objectives which in this work is to improve the global solutions.

2 Problem Definition

In the current study agent-based framework (as shown in figure 1, section 4) incorporating IP search methodology that fulfills the requirements of zero hard constraints values and minimum values for soft constraints is proposed. The proposed framework is used to test on real-life datasets at UMSLIC as shown in table 1. The objective of the problem is to develop communication protocol that helps agents in the framework to share the best part of the solution, guide the agents towards more promising search space, and hence find the improved feasible timetable solutions. The problem involves several hard constraints and soft constraints. Hard constraints need to be satisfied in all circumstances, whereas soft constraint violations should be minimised to increase the timetable quality, and increasing the satisfaction of the people who are affected by the timetable. The constraints undertaken in this work are explained in the following subsections.

Essentially the problem in the current study involves allocating a set of 35 timeslots (seven days, with five fixed timeslot per day) according to UMSLIC teaching guidelines. Each lecturer teaching several courses in each semester and each course has at least one lecture of minimum two hours per week. In addition, UMSLIC's administration has a guideline as shown in table 2 for the compulsory, elective, center for promotion of knowledge and language learning (PPIB), and center for co-curriculum and student development (PKPP) courses to be enrolled by the students in each of the semesters throughout the students' university days

Our approach will consider certain lecturer's preferences, better utilization of appropriate room and improved evenly student's schedule. Moreover, our approach also fulfill university teaching guideline where there are some general preferences such as some courses particularly program and faculty courses cannot be scheduled on weekends and must be scheduled on the first or third timeslots of the weekdays. In addition, some course such as PKPP courses cannot take place on weekdays. In addition, some courses such as PPIB course must be scheduled on second, fourth, or fifth in timeslot.

Hence, this research concentrates on real-life CB-UCT. In fact, in CB-UCT there are five variables identified namely periods, courses, lecturers, rooms, and curricula. The objective is to assign a period and a room to all lectures of each course according to the hard and soft constraints based on UMSLIC teaching guidelines. This research work aims to implement agent-based incorporating an IP search methodology for solving real-life CB-UCT for UMSLIC.

Table 1. Summary of the dataset from UMSICL academic division

	Semester1 s2016/2017	Semester2 2016/2017
Number of student	2263	2224
Number of curriculum	65	49
Number of lectures	108	92
Number of courses	134	117

2.1 Hard Constraints

Listed below are all the predefined hard constraints considered in this work:

1. *Lectures.* Each course has a predetermined amount of lectures that must be given. Each lecture must be scheduled in distinct time slots and the total number of lectures cannot be exceeded.
2. *Room conflict.* Each Two lectures cannot take place in the same room in the same time slot.

3. *Main and PPIB courses.* All main (major) courses cannot be scheduled at weekend. This is according to UMSLIC teaching guideline. Main course involves program core and school course as well as some PPIB courses.
4. *Center for co-curriculum and student development (PKPP) courses.* All PKPP courses must be scheduled at weekend. There are some courses under PKPP which by default must be scheduled at weekend. Normally this course is taught at the early semesters of the students' university years.
5. *Room Capacity.* The size of the room must be larger or equal to the size of the course. The room where the course is scheduled should be large enough to accommodate the number of students registered for that course.
6. *Curriculum and lecturer conflicts.* Lectures of courses in the same curriculum or taught by the same lecturer must all be scheduled in different time slots

2.2 Soft Constraints

Listed below are all the predefined soft constraints considered in this work:

1. *Lecturer preferences.* The assignment of classrooms and periods of time must allow satisfying at best the preferences of lecturers. I.e. there should be a gap for lectures taught by same lecture as well as the lecturers can specify times when they prefer not to lecture.
2. *Appropriate room size.* The usage of appropriate room size i.e. does not schedule a lecture with 30 students in a room with capacity of 300 seats.
3. *Evenly timetable.* The Student should not have consecutive courses per any given day.

Table 2. UMSKAL teaching guideline. Where 1 stands for Faculty courses, Program courses or elective courses; 2 stand for PPIB courses; 3 stand for PKPP courses.

Day/Time	**Time groups**	
	Monday - Friday	Saturday & Sunday
08.00 AM – 10.00 AM	1	3
10.00 AM – 12.00 PM	2	3
02.00 PM – 04.00 PM	1	3
05.00 PM – 07.00 PM	2	3
07.00 PM – 10.00 PM	2	3

3 Related Works

In general, there are many techniques proposed in literatures for solving timetabling problems in particular curriculum-based course university timetabling. However, scholars in operational research and artificial intelligence acknowledge Metaheuristics as indispensable techniques to address difficult problems in numerous and diverse fields [5]. Likewise, recently hype-heuristics has been widely used to address the issues. Nevertheless, even meta-heuristics may reach quite rapidly the limits of what may be addressed in acceptable computing times for many problem settings for research and practice alike [6, 18]. Similarly, hyper-heuristics do not generally guarantee optimality, performance often depending on the particular problem setting and instance characteristics [7]. Therefore, this thought has led birth of the fertile field of cooperative search especially in the operational research and artificial intelligence research community.

Generally, cooperative search can be natural approach to address the issues resulted from meta-heuristics and heuristics alike. [16] Stated that, "instead of trying to design new algorithms without downside, a task that is quite difficulty if not impossible, scholars in operational research and artificial intelligent research community have been working on the ways to organize the existing techniques in order to suppress their weakness through cooperation, and together do what separately they might not be able to accomplish". Ultimately parallel implementations of sequential algorithms appear quite naturally as an effective alternative to speed up the search for approximate solutions of combinatorial optimization problems [8]. Moreover parallel implementations allow solving larger problems or finding improved solutions, with respect to their sequential counterparts, due to the partitioning of the search space and to more possibilities for search intensification and diversification [4, 8, 19].

However, even with recent enormous effort in cooperative search, [15] believes this area has been little explored in operational research. Also, as computers keep becoming very powerful nowadays, this present huge opportunity for researcher to do what was unable to be done in 20 to 30 years ago especially in parallel computational research area. Similarly, according to [9] in recent years, multi-core processors are widely used and cooperative search can easily benefit from parallel processing. Thus in the last few years research community have started to exploit the opportunity presented by multi-core processors and work on how to develop optimization technique that is faster, more robust and easier to maintain.

More research in combinatorial optimization is currently being devoted in cooperative search techniques. Several number of cooperative search approaches have been proposed in the literature [4, 10]. The key idea behind cooperative search is to combine the strengths of different (meta-) heuristics to balance intensification and diversification and direct the search towards promising regions of the search space [4, 11]. Essentially, by cooperating the chances of finding novel and greatly improved solutions are increased.

[12] Defined cooperative search as the "parallelization strategy for search algorithms where parallelism is obtained by concurrently executing several search programs". In general cooperation by these programs is to interact with one another directly or indirectly, synchronously or asynchronously. Therefore the communication and sharing of information is an important feature of cooperation in cooperative search field [15]. The need to interact in such systems occurs because programs (agents) solve sub problems that are interdependent, either through contention for resources or through relationships among the sub problems [13]. The benefit of this

approach is the fact that, it adds parallel computational resources and possibility of information exchange (exchange best part of the solution) among the agents [14]

However, so far most of cooperative search focus more on metaheuristics and heuristics. Interestingly, integer programming search naturally offers significant opportunities for parallel computing, yet not enough research has been devoted to parallel integer programming implementations. In this research, we propose asynchronous agent-based framework incorporating integer programming search methodology for solving real-life CB-UCT at UMSLIC.

4 Agent-Based CB-UCT IP Framework

Figure 1 present the proposed agent-based searches framework. In this research, a decentralized agent-based framework, which consist of given number of agents (n) is proposed. Basically a framework is a generic communication protocol for integer programming (IP) search methodology to share solutions among each other. Each IP is an autonomous agent with its own representation of the search environment. To this end they share complete feasible solution to enable each other to direct (move) towards more promising search space. Moreover, the communication or ability for the agent to exchange the solutions with one another via the central agents prevent individual agent from stacking on the local optima [8]. Essentially all agents in the distributed environment communicate asynchronously via the central agent. Additionally, it is worth mentioning that, the initial feasible solution is generated by the central agents as well. In clarity this framework will involve asynchronous cooperative communication as follow.

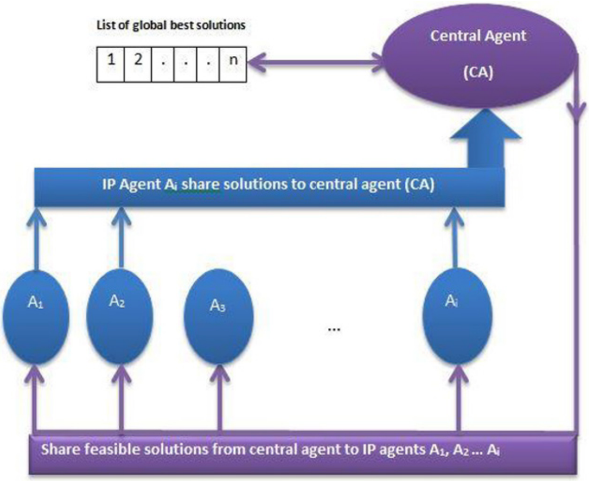

List of global best solutions
1
2
.
.
.
n
Central Agent
(CA)
IP Agent A_i share solutions to central agent (CA)
A_1
A_2
A_3
...
A_i
Share feasible solutions from central agent to IP agents A_1, A_2 ... A_i

Fig. 1. Proposed Agent-based IP Search methodology Framework

4.1 Central agent (CA)

The central agent is responsible to generate the initial feasible solutions as well as to coordinates the communication process of all other agents involved in the proposed framework. The central agent acts as intermediate agent among other agents where it passes the feasible solution and other parameters to the IP agents asynchronously on top of FIPA-ACL communication protocol. On top of that, the central agent receives the improved solution from the IP agents and compares the objective function cost value of the received solution with the existing global solutions on the list, if the improved solution's objective is better or similar to any of the solutions on the existing solutions then the worse in the list is replaced. Else the received solution is discarded and the central agent randomly select other solution from the list of the global solutions and send back to that particular agent so in order for the agent to try to improve the new solution received from the central agents.

4.2 IP Agents (A_i)

All other agents' start from the complete solution received randomly from central agent and iteratively perform search to improve the solution autonomously (independently). In this case the agents have to maintain the feasibility of the solution i.e. do not violent hard constraint. After certain number of iterations according to the rules stated (after every 10 seconds and no improvement found) the agent passes the solution back to the central agent and request new solution from the central agent. The central agent accepts the solution if only the solution is better or similar to the existing global solutions in the list of the solutions else the solution is discarded. If the solution is accepted then the solution with higher objective cost function i.e. worse in the list will be replaced. The reason an IP agent's exchange solution is to make sure the agents are not stuck on local optima, moreover scholars highlighted on the literature that, by exchanging the solution the possibility of the agents (algorithms) changing the position towards more promising search space is increased [4, 8, 19].

Best solution Criteria. All of our agents are incorporated with integer programming search methodology. Each agent also is capable to compute the final objective function and return it along with the improved solution. The central agent places all the solutions obtained in a sorted list where the solution on top will be the best solution (the solution with minimum objective function value).

In this framework the value of the objective functions is used to determine the quality of the solution. The lower the cost value the better the solution. Hence for the solution which has improved by the IP agents to be considered better than or similar to the global existing solutions, the returned improved solution's objective function should be lower than or similar to the one of the available in the global solutions objective functions values. Else the solution is discarded. The objective is to enable the IP agent to escape from local optimal and more importantly to allow the agent to move towards the most promising search space by sharing the best part of solution.

The whole process stops when all the IP agents are not improving the solution any more in a given number of conversations. Conversation in this regards means number of communication between the central agent and improving. For example IP agent A_i request new solution from central agent try to improve the solution however the agent is unable to improve anymore for three consecutive conversations. In this case the agent has reach appoint where unable to improve the solution anymore.

Proposed Agent-framework's Commitments rules.

The communication of an agent is built on top of FIPA-ACL protocol. The send and receive massage mechanism is well explained in the subsequent sub-sections pseudocode. The agents are in the agent society so each agent in the pack of agents follows the following commitments rules explained in as follow.

```
Commitments Rules (Pseudocode).
Let
Central agent is denoted as CA,
IP agents are denoted as Ai.

  {CA, REQUEST, DO (time, action)
          },;;; msg condition
          (B,
  [Now, Friend agent] AND
  CAN (self, action) AND
  NOT [time, CMT (self, anyaction)
  ),;;; mental condition
  DO (time, self, action)
  }
```

The proposed framework's commitments rules pseudocode may be paraphrased as follows:
If IP Agent *(A_i)* receives a message from central agent (*CA*) which requests A_i to do action (improve the solution) at time t, and *Ai* believe that; *CA* is currently a friend; and A_i can do the action; at time t, and A_i not committed to doing any other action, then A_i will commit to doing that action at time *t*. All agents in the framework are following this set of rules. These set of rules, guide agent in the framework on what to do on a given time to make sure agents do not interfere one action with another

5 Experimental Setup and Results

Now we discuss the performance of the proposed agent-based framework for CB-UCT, in which two-semester problem instances of different difficulty is tackled. For each semester (session one (s1) 2016/2017 and session two (s2) 2016/2017) datasets, the initial solutions generated by the central agent using pure 0-1 IP. In average the initial solutions are generated in five seconds. To determine the consistence of the al proposed framework, for each instance, we run the experiments 50 times and the average final costs are computed in table 3. In this experiment, first we use three IP agents (Ai), and then we increase the number of IP agents (Ai) from three to six IP agents (Ai).

The improvement from initial to final cost value when three IP agents (Ai) are used is 12.73% and 10.20% for s1 2016/2017 and s2 2016/2017 respectively. On the other hand, the improvement of the solution's cost value when six IP agents (Ai) is used are 17.89% and 15.58% % for s1 2016/2017 and s2 2016/2017 respectively. The main benefits of the agent-based approach adopted for CB-UCT are the possibilities of intensifying and diversifying the search space, where Ai is able to changes solutions among each other in the distributed MAS. This leads the IP agents to easily move towards the most promising search areas of the search space. Basically, by the analysis the results, the numbers of IP agents used in the framework determine the quality of the solution generated. In this regard we find out the quality of the solution in this framework proves to increase slightly as the number of IP agents (Ai) are increased

Table 3. Experimental results for the proposed agent-based search framework.

	No of agents	Semester1 s2016/2017	Semester2 S2016/2017
Initial cost	-	368.04	377.29
Final average cost	3	321.20	338.80
Final average cost	6	302.20	318.50
Average improvements (%)	3	12.73	10.20
Average improvements (%)	6	17.89	15.58

6 Conclusion and Future Work

The current study focuses on agent-based IP framework for the CB-UTT for real-life instances in UMSLIC. The proposed framework is able to produce an applicable solution for UMSLIC. Based on the methodology employed, it is discovered that the sharing of solutions among agent improved the overall performance of the framework as the number of agent increase the solution quality slightly improve.

The currents study recommends that the future work may include agent negotiation; the negotiation amongst the IP agents (A_i) may lead to better performances of the proposed agent-based search framework.

References

1. Oprea M.: Multi-Agent System for University Course Timetable Scheduling. The 1st International Conference on Virtual Learning, ICVL (2006)

2. Babaei, H., Karimpour, J., & Hadidi, A. (2015). A survey of approaches for university course timetabling problem. Computers & Industrial Engineering, 86, 43-59.
3. Obit. J. H., Ouelhadj, D., Landa-Silva, D., Vun, T. K.., Alfred, R.: Designing a multi-agent approach system for distributed course timetabling. IEEE Hybrid Intelligent Systems (HIS), 10.1109/HIS(2011)-6122088.
4. Obit, J. H., Alfred. R., Abdalla, M.H.: A PSO Inspired Asynchronous Cooperative Distributed Hyper-Heuristic for Course Timetabling Problems. Advanced Science Letters, (2017)11016-11022(7)
5. Crainic, T. G., Toulouse, M.: Parallel strategies for meta-heuristics. In Handbook of metaheuristics (pp. 475-513): Springer (2003).
6. Blum, C., Puchinger, J., Raidl, G. R., & Roli, A. (2011). Hybrid metaheuristics in combinatorial optimization: A survey. Applied Soft Computing, 11(6), 4135-4151.
7. Crainic, T.G.: "Parallel meta-heuristic search", Tech. Rep. CIRRELT-2015-42, (2015)
8. Cung, V.-D., Martins, S. L., Ribeiro, C. C., Roucairol, C.: Strategies for the parallel implementation of metaheuristics. In Essays and surveys in metaheuristics (pp. 263-308): Springer (2002)..
9. Yasuhara, M., Miyamoto, T., Mori, K., Kitamura, S., Izui, Y.: Multi-objective embarrassingly parallel search for constraint programming. Paper presented at the Industrial Engineering and Engineering Management (IEEM), 2015 IEEE International Conference
10. Crainic, T. G., Gendreau, M.: A Cooperative Parallel Tabu Search for Capacitated Network Design, Technical Report CRT-97-27(1997).
11. Ouelhadj, D., Petrovic, S.: A cooperative hyper-heuristic search framework. Journal of Heuristics, 16(6) (2010)., 835-857.
12. Toulouse, M., Thulasiraman, K., & Glover, F. (1999, August). Multi-level cooperative search: A new paradigm for combinatorial optimization and an application to graph partitioning. In European Conference on Parallel Processing (pp. 533-542). Springer, Berlin, Heidelberg.
13. Lesser, V. R.: Cooperative multi-agent systems: A personal view of the state of the art. IEEE Transactions on knowledge and data engineering, 11(1) (1999), 133-142.
14. Silva, M. A. L., de Souza, S. R., de Oliveira, S. M., & Souza, M. J. F.: An agent-based metaheuristic approach applied to the vehicle routing problem with time-windows. Paper presented at the Proc. of the Brazilian (2014) Conference on Intelligent Systems-Enc. Nac. de Inteligência Artificial e Computacional (BRACIS-ENIAC 2014).
15. Martin, S., Ouelhadj, D., Smet, P., Berghe, G.V., Özcan, E.: Cooperative search for fair nurse rosters. Expert Syst. Appl. 40(16), 6674–6683 (2013).
16. Talukdar, S., Baeretzen, L., Gove, A., and de Souza, P.: Asynchronous teams: Cooperation schemes for autonomous agents. Journal of Heuristics, 4:295–321, 1998.
17. Wooldridge, M, Jennings, N.: Intelligent Agents, Lecture Notes in Artificial Intelligence 890 Springer-Verlag (eds.), 1995b.
18. Obit. J. H, Landa-Silva, D.: Computational Study of Nonlinear Great Deluge for University Course Timetabling, Intelligent Systems - From Theory to Practice, Studies in Computational Intelligence, Vol. 299, V Eds. Springer-Verlag, 2010, pp. 309-328
19. Obit, J. H.: Developing novel meta-heuristic, hyper-heuristic and cooperative search for course timetabling problems. Ph.D. Thesis, School of Computer Science University of Nottingham (2010).

3D Face Recognition using Kernel-based PCA Approach

Marcella Peter, Jacey-Lynn Minoi and Irwandi Hipni Mohamad Hipiny

Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, Malaysia.
marcellapeter@yahoo.com

Abstract. Face recognition is commonly used for biometric security purposes in video surveillance and user authentications. The nature of face exhibits non-linear shapes due to appearance deformations, and face variations presented by facial expressions. Recognizing faces reliably across changes in facial expression has proved to be a more difficult problem leading to low recognition rates in many face recognition experiments. This is mainly due to the tens degree-of-freedom in a non-linear space. Recently, non-linear PCA has been revived as it posed a significant advantage for data representation in high dimensionality space. In this paper, we experimented the use of non-linear kernel approach in 3D face recognition and the results of the recognition rates have shown that the kernel method outperformed the standard PCA.

Keywords: 3D face, Facial recognition, Kernel PCA.

1 Introduction

Face recognition is a biometric system aim to identify a person in a digital image by analyzing and comparing facial patterns. Face recognition process may be a simple task for human but it is a challenge and difficult task for the computer as it requires the involvement of a combination of statistical approaches, Artificial Intelligent, computer vision and machine learning methods. As mentioned in [1], face recognition rates dropped when facial expression is included in a system. In addition, other facial variants and factors such as pose, facial expression, hairstyle, makeup, mustache, beard and wearing glasses would also contribute to the low recognition rate [2].

For the past two decades, various face recognition techniques were developed by researchers with the aim to improve the performance of face recognition. The general pipeline for an automated face recognition system involves four steps: facial image acquisition and detection, face feature extraction, and face classification and recognition [3, 4]. Facial image acquisition and detection are usually done in the pre-processing phase. This is followed by facial feature extraction process, where features such as the eyes, the nose, and the mouth are extracted and classified. Facial features are important in many face-related applications, as in face alignment and feature extraction. The conventional method in feature extraction commonly used linear transformation approach, i.e. the Principal Component Analysis (PCA). PCA is a common

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_8

and yet powerful dimensionality reduction technique for extracting and projecting data structures into lower subspaces. It is readily performed by solving eigenvalues and eigenvectors to extract discriminant principal components [6].

1.1 From PCA to Kernel-based PCA

PCA is an effective technique commonly used in face recognition study for discriminant analysis. PCA is widely implemented in 2D and 3D face recognition systems as the base method [1, 17, 18]. Depth information in 3D data makes face recognition system to become more robust because it does not depend on illumination and pose [3]. Recent work by [1] implements 3D model-based face recognition system that also utilizes PCA method to recognize person under any facial expressions. However, the limitation is that PCA could not deal with large range of face variants. In other words, to simplify object with complex structures in a linear subspace, PCA might not be sufficient enough [6]. 3D face data has a complicated structure in a high dimensional space. The nature of non-linear shape exists when the face images are projected under different conditions with different lighting, expression variant, identity variant, and poses that lead to shape deformations [7, 8]. Due to this matter, a kernel approach is proposed to overcome the non-linearity. The idea of kernel is to map the input face images into a higher dimensional space in which the manifold of the face linearity is simplified.

In comparison to other non-linear techniques of feature extraction, kernel method does not require non-linear optimization, and only the solution from eigenvalue measurements [9]. Kernel learns non-linear functions while working with several training examples. Kernel can make the data linearly separable by projecting the data onto a higher dimensional feature space and we use a kernel function to do this. By choosing different form of kernel function, it can handle different non-linear problems. Kernel functions that most researchers adopt are the Polynomial kernel and Gaussian kernel. Kernels have been successfully used in Support Vector Machine (SVM) for classification purpose. One of the reason for SVM [10] success is the 'kernel trick' implicitly reduce computations of data in feature space of each input vector and it directly gives result into the feature space [9, 13].

Schölkopf et al. [11] extended classical linear PCA to Kernel PCA method. Kernel PCA projects the input space into feature space, by mapping the principal component vectors using kernel function. Recent work by Wang [12] has applied Kernel PCA to explore the complicated structure of 2D image for face recognition. Pre-image is reconstructed using Gaussian kernel PCA and then use it to design Kernel-based Active Shape Model [12]. Their results have shown that the error rate is lower with 11.54% for Kernel PCA compared to linear PCA with 23.06%.

2 Related Works

Researches in 3D face recognition are mostly using linear approaches and techniques [1, 17, 18]. There also exist non-linear approaches and they are being used and improved for face applications. Recent works by [19] used discriminative depth estimation approach that adopts convolutional neural network, which retains the subjects' discriminant information and multilayer convolutional network, to estimate the depth information from a 2D face image. Their approach investigates the manifold of network layers from a single color image to extract new space, where in this case the depth space, for discriminant features to be extracted using the network to improve the 2D face recognition.

A similar non-linear approach by [20] proposed a 3D morphable model that used three deep neural networks aim to reconstruct the 3D face data input. The proposed fitting algorithm works by estimating network encoder from decoded two parameters shape and texture, with the assistance of a geometry-based rendering layer. The proposed works are directed towards applications in 3D facial recognition and facial synthesis. However, these two approaches require a large number of parameters which may lead to overfitting of training data when implemented into a small number of data. Therefore optimization such as using non-linear regression approximation [21] is needed to resolve the problem. The idea to extract hidden layers without high computation and taking the advantage of kernel methods as indicated in [9] motivates this research to explore Kernel PCA method to extract non-linear principle components from a 3D data, which is more practical to work on with a small number of dataset.

This project will use existing 3D face dataset. The project hence move on to the analysis of a statistical model that can fit 3D face data points as input data to define a shape space by selecting basis direction that holds the greatest variance of a face data where covariance matrix is then computed for the face data to be projected into the shape space. The paper will present the non-linear Kernel PCA approach on 3D face datasets and its experimental analysis by comparing the recognition rates with the baseline.

3 Kernel-based PCA

The algorithm of Kernel-based PCA is adopted and can be found in [14]. Given a set of m centered or with zero mean samples

$$x_k = [x_k, \dots, x_{kn}]^T \in R^n, \tag{1}$$

The purpose of PCA is to find the directions of projection that get the most out of the variance C, which is corresponding to finding the eigenvalues from the covariance matrix

$$\lambda w = Cw \tag{2}$$

for eigenvalues $\lambda \geq 0$ and eigenvectors $w \in R^n$. In Kernel PCA, each vector x is projected from the input space, R^n, to a high dimensional feature space, R^f, by a nonlinear mapping function:

$$\Phi: R^n \to R^f, f \gg n . \tag{3}$$

Note that the dimensionality of the feature space can be huge. In R^f, the corresponding eigenvalue problem is $\lambda w^{\Phi} = C^{\Phi} w^{\Phi}$ where $C^{\Phi}$is a covariance matrix. All solution w^{Φ} with $\lambda \neq 0$ lie in the span of $\Phi, \ldots, \Phi(x_m)$ and there exist coefficients $a_i$such that

$$w^{\Phi} = \sum_{i=1}^{m} a_i \Phi(x_i) \tag{4}$$

Denote an $m \times m$ matrix K by

$$K_{ij} = K(\Phi_i, \Phi_j) = (\Phi_i).(\Phi_j), \tag{5}$$

the Kernel PCA problem becomes

$$m\lambda K a = K^2 a \tag{6}$$

$$m\lambda a = K a \tag{7}$$

Where a denotes a column vector with entries $a_1, \ldots, a_m$. The above derivation assumes that all the projected samples $\Phi(\mathrm{x})$ are centered in R^f . As to centralize the vectors $\Phi(\mathrm{x})$ in R^fcan be found in [9]. We can now project the vectors in R^f to a lower dimensional space spanned by the eigenvectors w^{Φ}. Let x be a test sample whose projection is $\Phi(\mathrm{x})$ onto w^{Φ}, is the nonlinear principal components corresponding to

$$\Phi: w^{\Phi}.\Phi(\mathrm{x}) = \sum_{i=1}^{m} a_i\big(\Phi(x_i) \,.\, \Phi(x_i)\big) = \sum_{i=1}^{m} a_i\,(x_i, x) \tag{8}$$

The first q $(1 \leq q \leq m)$ nonlinear principle components or the eigenvectors w^{Φ} are extracted using the kernel function without expensive operation that explicitly projects samples to high dimensional space R^f. The first q components correspond to recognition where each x encodes a face image.

According to [15], first order polynomial kernel of Kernel PCA is a special case of traditional PCA. The polynomial kernel can be expressed as

$$K(x, y) = (x^T y)^d \tag{9}$$

where, the power of d is specified or formed beforehand by the user.

4 Experiment and Results

An experiment is conducted using Kernel PCA on 3D face dataset to demonstrate its effectiveness and compared results obtained with linear PCA. The Kernel PCA is used as feature extraction through matrix decomposition to extract nonlinear principal components (eigenvectors) and Kth Nearest Neighbor classifier is used to measure the Euclidean distance as classification mechanism.

4.1 Datasets

The experiment is carried on Imperial College London 3D face database that contains 240 face surface models of 60 subjects with 4 females and 56 males. The subjects were also classified in terms of their ethnicity as 8 South Asians, 6 East Asians, 1 Afro-Caribbean and 45 Caucasians. These subjects were mostly students within an age range of 18-35 years. The facial, and was acquired in several different head positions and three facial expression poses. The facial expressions were smiling, frowning, and neutral. The 3D face dataset has already been preprocessed, thus preprocessing stage is omitted. Details of the preprocessing can be found in [22]. Fig. 1 shows the example images adopted from Imperial College London in 2D (left) and 3D (right) environment.

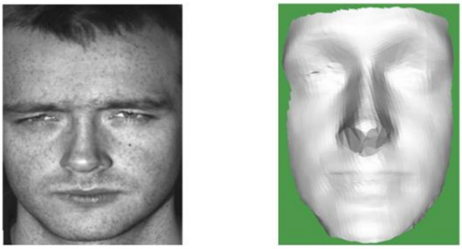


Fig. 1. Sample from Imperial College London face database [15]

4.2 Relationship between distance measurement to face recognition result

The PCA face recognition system is normally implemented along with Kth Nearest Neighbor (KNN) algorithm. The algorithm finds the closest K neighbors with minimum distance from a subspace to classify a testing set. Different parameters can be used with KNN such as different value of K and distance model. The K value in KNN during the recognition process affects the overall face recognition rate. As the recognition process is based on the shortest distance, therefore the face recognition rate varies based on different type of distance classifier such as Euclidean distance and Manhattan distance which are used widely in evaluating the rate of facial recognition. In this project, Euclidean distance is chosen simply because of the idea to measure the shortest distance, for example the length between the straight lines of two points. In general, the distance between two points of x and y, in Euclidean space, R^n is given by:

$$d(x,y) = \|x-y\| = \sqrt{(|x_i - y_i|)^2} \tag{10}$$

4.3 Experimental Plan

The 3D face recognition experiment started with training and followed by recognition process. Herewith, a proposed experiment procedure by using cross validation approach. The experiment begins with two of the face class generated from 3D face dataset are selected and used as a training set and testing set for face recognition. As shown in Table 1, there are four sets of 3D face database that are classified according to different expressions. The four classes are named as neutral 1, neutral 2, frowning and smiling, with 60 subjects for each set. As mentioned in [16], the use of different K value for different instances may give positive outcome for classification accuracy. Thus, we followed the approach proposed in [16] which allows selection of a local value of K to find the best classification. In this experiment, the K value = 3 is used, which yields the lowest number of error during classification. Equation (11) is used to evaluate the rate of recognized face for each testing set.

$$Face\ recogntion\ rate = \frac{Correct\ Match\ Count}{Total\ face\ test\ count} \times 100\ \% \tag{11}$$

Table 1. Number of samples of each face dataset.

Face class	Number of sample
Neutral 1	60
Neutral 2	60
Frowning	60
Smiling	60

Training. The face recognition system read each subject from the training dataset selected and extract 3D points based on their referred directory path. Next, each training subject with their label using Kernel PCA is projected into the training subspace, which acts as a knowledge base for the recognition process. This process is iterated until all 60 subjects are projected into the training subspace.

Recognition. The face recognition system read each subject from the 3D testing dataset selected and extract 3D points based on their referred directory path. Next, each testing subject is projected into the training subspace using Kernel PCA. From the feature vectors obtained from the subspace projection, determine Euclidean distance for each test subject and the training subjects based on the selected K value (3 NN). After that, the distance between the testing subjects with the training subjects are compared. Then, the testing subject is classified as the same label as the training subject that has the shortest distance. The process is iterated until all 60 subjects in testing dataset are tested. To validate the newly classified subjects, cross validation is performed and the rate of recognition is computed. Then, the result is displayed with the recognition rate computed earlier. The test is repeated by using different testing sets.

The experiments are arranged as Test 1: Frowning, Test 2: Neutral 1, Test 3: Neutral 2, and Test 4: Smiling. The following Table 2 presents the experiment result for Test 1. The results of face recognition using Kernel PCA are compared with PCA results, which has been set as the baseline approach in this research.

Table 2. Face recognition rate for Test 1 (%).

Test 1: Frowning (3-NN)		
Training Set	**Kernel PCA**	**PCA**
Frowning	98.33	100.00
Neutral 1	90.00	85.00
Neutral 2	83.33	76.67
Smiling	36.67	20.00
Total average rate:	**77.08**	**70.42**

4.4 Results and Analysis

Table 3 presents the average rate of 3D face recognition based on Kernel PCA and PCA, respectively. Based on the results, we found that by using Neutral 1 and Neutral 2 as the training set, both results would give the highest recognition as compared to Frowning and Smiling. Meanwhile for comparison purposed between Kernel PCA and PCA method, the average recognition rate shows that Kernel PCA achieved higher recognition rate than PCA.

Using frowning test set in Kernel PCA earns a recognition rate of 77.08%, while 70.42% of recognition rate was achieved when using PCA. Besides that, Neutral 1 test set in Kernel PCA could achieved a recognition rate of 82.50% meanwhile with PCA only earns 78.33% of recognition rate. Using Neutral 2 test set, Kernel PCA achieved a recognition rate of 85%, which is much higher than PCA that earns 77.92% of recognition rate. In the last experiment of the test set, smiling test set have shown 64.59% of recognition rate using Kernel PCA and PCA achieved 47.08% of the recognition rate. Fig. 2 illustrates the comparison of both methods. The overall recognition rate for Kernel PCA is 77.29%, which is higher than PCA with 52.69%.

Table 3. Average face recognition rate of two methods (%)

	Frowning	Neutral 1	Neutral 2	Smiling
Kernel PCA	77.08	82.50	85.00	64.59
PCA	70.42	78.33	77.92	47.08

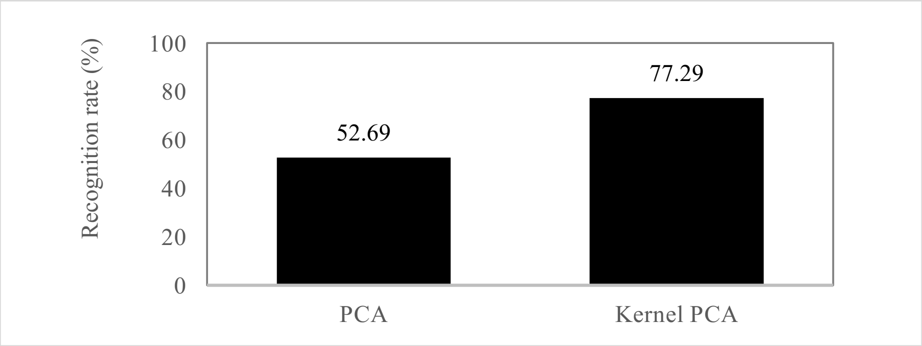


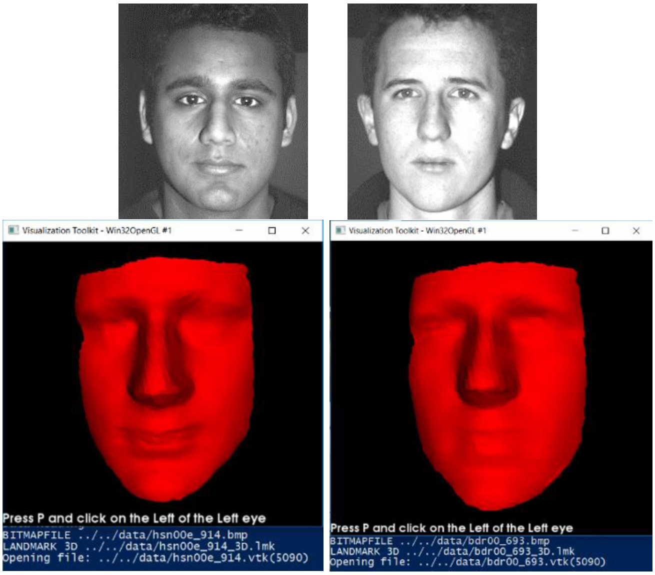
Fig. 2. Recognition rate of PCA compared to Kernel PCA (%).

4.5 Discussion

The comparison between Kernel PCA and PCA methods is as presented in Fig. 2 shows that Kernel PCA could achieved a higher recognition rate than PCA technique alone. We have also found out that training the neutral face will always provide higher face recognition rate if compared to using frowning and smiling faces. This is because neutral faces have lesser face variant compared to frowning and smiling. Smiling has more facial variances at the lower part which involve cheek and mouth. Meanwhile, frowning has more variances over top part of a face namely, the eyes and eyebrows. Therefore, in most of literature, neutral faces are preferably used for its lesser face variants.

From the experiment, we have also identified two misclassified faces using PCA. The two subjects (see Fig. 3) are incorrectly matched when Neutral 1 training sample is used on Neutral 2 testing sample. Based on the observation made in Fig. 3, the 2D face on the top left and 3D face at the bottom left is the same person but the recognition system has wrongly matched the face to the one on the right side. This could be due to the similarity of facial structures between the two subjects. However, when tested using Kernel PCA, the two faces were correctly matched. This shows that Kernel PCA could correctly extract facial features. However, in shown Table 2, when Frowning test set is tested with Frowning training set using Kernel PCA method, the recognition rate recorded at 98.33%. Based on the analysis, the test subject is correctly classified but with false rejection. This limitation problem will be further investigated in the future work.

The experiment has proved that Kernel PCA as a non-linear approach, able to extract the non-linearity property within face dataset to yield a higher face recognition rate compared to PCA. Noticed that Kernel PCA had increased the face recognition rate to 24.6% compared to PCA (see Fig. 2).



Visualization Toolkit - Win32OpenGL #1
Visualization Toolkit - Win32OpenGL #1

```
Press P and click on the Left of the Left eye
BITMAPFILE ../../data/hsn00e_914.bmp
LANDMARK 3D ../../data/hsn00e_914_3D.lmk
Opening file: ../../data/hsn00e_914.vtk(5090)
```

```
Press P and click on the Left of the Left eye
BITMAPFILE ../../data/bdr00_693.bmp
LANDMARK 3D ../../data/bdr00_693_3D.lmk
Opening file: ../../data/bdr00_693.vtk(5090)
```

Fig. 3. Incorrect match of two subjects.

5 Conclusion and Future Works

This paper described the application of Kernel-based PCA approach in face recognition domain. Kernel is an inherently multi-disciplinary domain. It is vital to look at it from all fields and perspectives to have an insight on how to develop an efficient automatic 3D face recognition system. The result of the overall recognition rate for Kernel PCA is 77.29%, while PCA has 52.69%. Experiment results proved that Kernel-based PCA approach outperforms linear PCA in 3D face recognition. The future work will focus on analyzing facial recognition method using other 3D face datasets, testing out other kernel methods and further investigate factors such as modifying the number of principal components and number of training samples.

Acknowledgement. The authors would like to thank Suriani Ab Rahman, Phoon Jai Hui for contributing to the research in this paper, and Newton Fund for the financial support for the publication.

References

1. Agianpuye, A. S., & Minoi, J. L.: Synthesizing neutral facial expression on 3D faces using Active Shape Models. In *Region 10 Symposium, 2014 IEEE* (pp. 600-605). IEEE. (2014).
2. Wen, Y., Lu, Y., Shi, P., & Wang, P. S.: Common Vector Based Face Recognition Algorithm. In *Pattern Recognition, Machine Intelligence and Biometrics* (pp. 335-360). Springer Berlin Heidelberg. (2011).

3. Hassabalah, M. & Aly, S.: Face Recognition: Challenges, Achievements and Future Directions. In IET Computer Vision Journals, Vol. 9, Iss. 4, pp. 614-626. (2015).
4. Chen, W., Yuen, P. C., Fang, B., & Wang, P. S.: Linear and Nonlinear Feature Extraction Approaches for Face Recognition. In *Pattern Recognition, Machine Intelligence and Biometrics* (pp. 485-514). Springer Berlin Heidelberg. (2011).
5. M. Kirby & L. Sirovich.: Application of the Karhunen-Lòeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(1):103-108. (1990).
6. Devi, R., B., Laishram, R., & Singh, Y., J.: Modelling Objects Using Kernel Principal Component Analysis. ADBU Journal of Engineering Technology 2.1. (2015).
7. Shah, J., H., et al.: A Survey: Linear and Nonlinear PCA Based Face Recognition Techniques. Int. Arab J. Inf. Technol. 10.6: 536-545. (2013).
8. Lee, C, S. & Elgammal, A.: Non-linear Factorized Dynamic Shape and Appearance Model for Facial Expression Analysis and Tracking. In IET Computer Vision, Vol. 6, Iss. 6, pp. 567-580. (2012).
9. Schölkopf, B., Smola, A., & Müller, K. R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, *10*(5), 1299-1319. (1998).
10. Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. & Vapnik. V.: Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Trans. Sign. Processing, 5:2758 –2765. (1997).
11. Schölkopf, B., Smola, A., & Müller, K. R.: Kernel principal component analysis. In *International Conference on Artificial Neural Networks* (pp. 583-588). Springer Berlin Heidelberg. (1997).
12. Wang, Q.: Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*. (2012).
13. Alaíz, C. M., Fanuel, M., & Suykens, J. A.: Convex Formulation for Kernel PCA and Its Use in Semisupervised Learning. *IEEE Transactions on Neural Networks and Learning Systems.* (2017).
14. Yang M. H.: Face Recognition Using Kernel Methods. Advances in Neural Information Processing Systems. MIT Press, 13: 960 – 966. (2001).
15. Imperial College London 3D face database. (n.d).
16. García-Pedrajas, N., del Castillo, J. A. R., & Cerruela-García, G.: A Proposal for Local k Values for k-Nearest Neighbor Rule. *IEEE transactions on neural networks and learning systems*, *28*(2), 470-475. (2017).
17. Okuwobi, I. P., Chen, Q., Niu, S., & Bekalo, L.: Three-dimensional (3D) facial recognition and prediction. Signal, Image and Video Processing, 10(6), 1151-1158. (2016).
18. Ouamane, A., Chouchane, A., Boutellaa, E., Belahcene, M., Bourennane, S., & Hadid, A.: Efficient tensor-based 2d+ 3d face verification. *IEEE Transactions on Information Forensics and Security*, 12(11), 2751-2762. (2017).
19. Cui, J., Zhang, H., Han, H., Shan, S., & Chen, X.: Improving 2D face recognition via discriminative face depth estimation. *Proc. ICB*, 1-8. (2018).
20. Tran, L., & Liu, X.: Nonlinear 3D Face Morphable Model. *arXiv preprint arXiv:1804.03786.* (2018).
21. Villarrubia, G., De Paz, J. F., Chamoso, P., & De la Prieta, F.: Artificial neural networks used in optimization problems. *Neurocomputing*, *272*, 10-16. (2018).
22. Papatheodorou, T.: 3d face recognition using rigid and non-rigid registration. PhD Thesis, Imperial College. (2006).

Mobile-Augmented Reality Framework For Students Self-Centred Learning In Higher Education Institutions

Aaron Frederick Bulagang and Aslina Baharum
Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
aaronfb91@gmail.com, aslinabaharum@gmail.com

Abstract. Augmented Reality is an exciting technology that enables user to view virtual object overlaying physical environment to increase the student-learning outcome. Objectives: To find if visualization plays a vital role mobile learning through comparison between three groups that were analysed using SmartPLS 3 using Measurement and Structural Model, Traditional, Mobile and mobile-Augmented Reality (mAR) Groups. The analysed data were collected at five universities. Methods: Quantitative method was conducted with students using questionnaire that covers Effectiveness, Self-Efficacy, Motivation, Satisfaction and Features. This paper also presents the comparison between the three groups in terms of Effectiveness towards several relationships that were analysed using SmartPLS to find significant relationship between the construct. It was found that Traditional had one significant relationship where as Non-mAR and mAR both have two significant relationships.

Keywords: Education; Framework; Higher Learning; Mobile-Augmented Reality

1 Introduction

This research main focus is implementing visualization tools such as 3D object, images and videos into higher learning education. The main method used here is mobile-Augmented Reality or mAR which overlays the content on top of physical object such as paper to show the visualization content to make the student be more interested learning. mAR is then compared to existing learning method such as the use of PowerPoint slide and mobile application. Jamali's research [1] was the main reference as it focuses mainly on mAR development for learning human anatomy using AR through HumAR.

2 Methodology

Throughout this research, there were five phases involved. The first phase of the research is to identify functional requirement, which is what functions are planned to be developed into the application. Current M-learning have complex UI [2] which is why multimodal interface was used as the main interface for the application that allows

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_9

multiple interactions such as touch manipulate and voice narration [7]. The second phase is to find the tools require developing the application such as smartphone, laptop and the software used such as Unity SDK, Vuforia and Blender [11]. Third phase is to start developing the prototype with the tools gather. The fourth phase is to test the developed prototype to find any error or bugs to fix. The fifth phase is to test the application with students to find if any external error or changes could be made into the application. The framework designed by Jamali [1] was adapted for this research as a guide and to compare with the result found through this research.

2.1 Participants and Procedure

The data collection was conducted at five Universities across Malaysia, which includes Universiti Pertahanan Negara Malaysia (UPNM), Management & Science University (MSU), Universiti Malaysia Sabah (UMS), and Universiti Teknikal Malaysia Melaka (UTeM) and Universiti Teknologi MARA (UiTM). All of the universities stated above offered Computer Science program that include Network Fundamental courses as the main subject used for this research as Programming and Computer organization application application has been developed by [3] and [4]. The research aims to find and analyse the best method for students to learn Network Fundamentals among three methods, which is Traditional via PowerPoint Slides, Non-mAR [10] or Mobile group utilize smartphone with an application installed and mAR group that utilize smartphone with an Augmented Reality application installed. Augmented Reality or AR is a technology that allows user to view virtual object in a physical environment [5], this makes the student feel more interested in learning with multimodal interface that includes 3D object manipulation, resizing and voice narration to allow easy navigation [8].

During the data collection at each university, the students were first brief about the research procedure. First, the student will be separated into three groups, which are Traditional, Non-mAR and mAR Group. Student in the Traditional group are iOS while Non-mAR and mAR group are majority Android User. This is because support for iOS had yet to be developed. The groupings for students are shown in Figure 1.

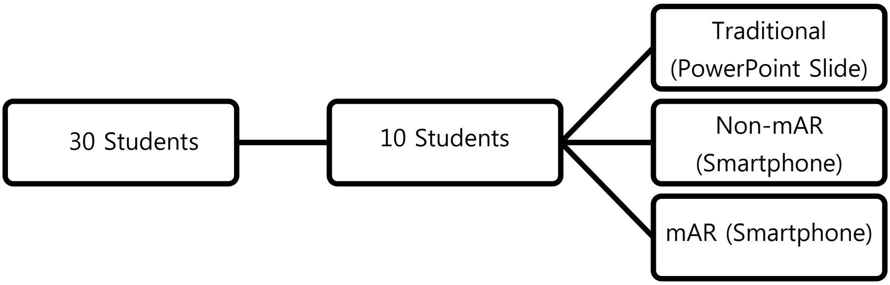


Fig. 1. Grouping of students during data collection

As shown in Figure 1, after grouped, each of the students will be taking a pre-test quiz which will question the student about their primary understanding about the 7

Open System Interconnect (OSI) Layer, Router and Switch. The students are then given their learning material for each group. The Traditional Group received a printed PowerPoint slide, The Non-mAR group smartphones were installed with an application to learn about Network Fundamentals while the mAR group smartphone are installed with NetmAR (Networking Fundamental mobile-Augmented Reality) application that allows the student to view the 3D models of the Router and Switch in real time with added information such as the port labels that exist. The procedure during data collection is shown in Figure 2.

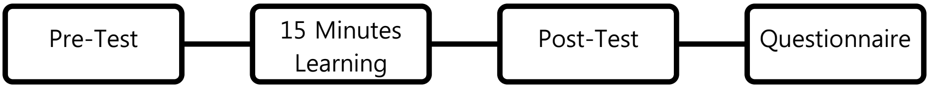


Fig. 2. Procedure during data collection

The students were given 15 minutes to study with their respective learning material according to their group. After the 15 minutes are up, they were given another quiz in the Post-Test session. In the Post-Test quiz, the question is relatively similar except for few minor changes such as the structure and the sequence of the question. Next, distribute the questionnaire to the students according to their group regarding their overall experience using their learning material. The questionnaire asked the student regarding the Effectiveness, Self-Efficacy, Motivation, Satisfaction [2] and Features of the learning material that they have used. After the data has been collected, the data are then entered into Statistical Package for the Social Sciences (SPSS) for normalization and to be exported into SmartPLS 3 for further Analysis. For SmartPLS 3, the data were analysed by grouped, first the Traditional Group are analysed through Measurement Model and Structural Model, followed by Non-mAR and mAR group.

The Measurement Model is to eliminate the items in the questionnaire that received a scoring below 0.6 [6], this is to ensure that the data produce is according to which stated that the item has to be above 0.6 to be considered acceptable. After the Measurement Model process is complete, the Structural Model can begin, the Structural Model is to analyse the data from the Measurement Model to investigate if the relationship between Satisfaction, Motivation, Self-Efficacy and Features are significant towards Effectiveness. It is done using two-tailed method with 500 subsamples.

3 Results and Discussion

3.1 Traditional Group

In the traditional group, overall there were 68 students that participated, where 51.5% (35 students) are female while 48% (33 students) are male. The following figures show the Traditional Method Measurement Model and Structural Model of the group. Figure 3 shows the Measurement Model, where the blue circle represents the Construct or the Main title of the items in the questionnaire, which consist of Self-Efficacy, Motivation and Satisfaction towards Effectiveness [2]. The yellow rectangle represents the items asked in the questionnaire and the remaining items that are above 0.6 [6].

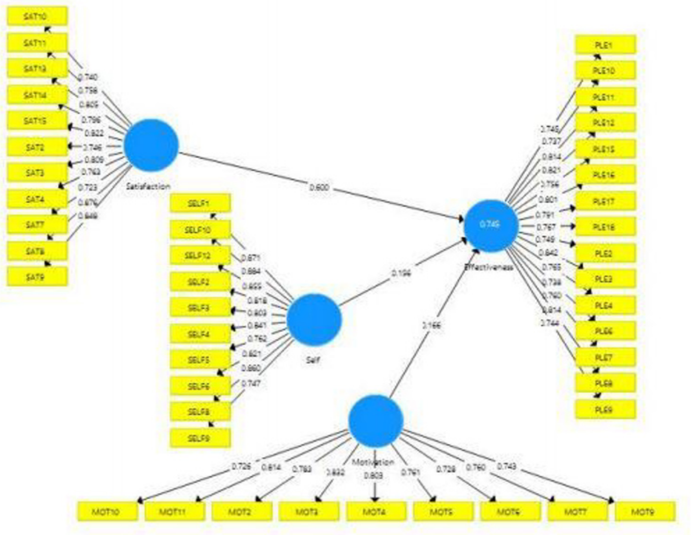


Fig. 3. Measurement Model

In Figure 4 shows the Structural Model for traditional group where the data from Measurement Model has been analysed and shows the significant relationship towards Effectiveness. In order for the relationship to be considered significant, it has to be above 1.65 of T-Value[a] to be considered significant. Table 1 shows the result of the significance testing, out of the three relationships; only one relationship is significant towards Effectiveness, which is Satisfaction with a T-Value[a] of 5.332, which exceeds the 1.65 with a significant level of 0.01. Overall, Traditional Method Group only has one significant relationship. This shows that the students are satisfied with their learning material but lack in Motivation and Self-Efficacy.

Table 1. Significance Testing Results of the Structural Model Path Coefficients

Relationship	**Original Sample (O)**	**Sample Mean (M)**	**Standard Deviation (STDEV)**	**T-Value[a]**	**Significant Level**
Motivation -> Effectiveness	0.166	0.158	0.121	1.373	Not Supported
Satisfaction -> Effectivness	0.600	0.614	0.112	5.332	0.01
Self -> Effectiveness	0.156	0.153	0.160	0.973	Not Supported

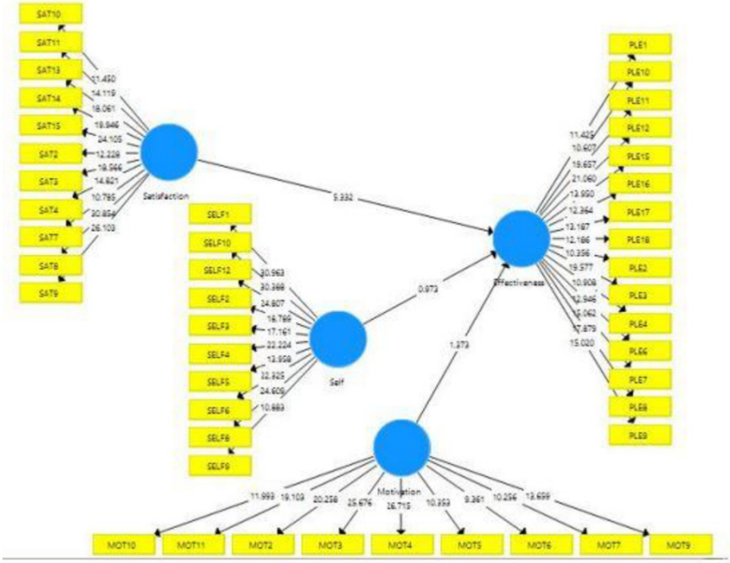


Fig. 4. Structural Model

3.2 Non-mAR Group

In the Non-mAR group, the students utilize an application developed using MIT App inventor for a simple and easy to use application to learn about Network Fundamental. Sixty-five students participated where 49.2% (32 Students) are female and 50.8% (33 Students) are male. Figure 5 shows the Measurement Model for Non-mAR Group where the blue circle represents the Construct, which consists of Motivation, Self-Efficacy and Satisfaction towards Effectiveness. The yellow rectangle represents the items in the questionnaire that is above 0.6 values.

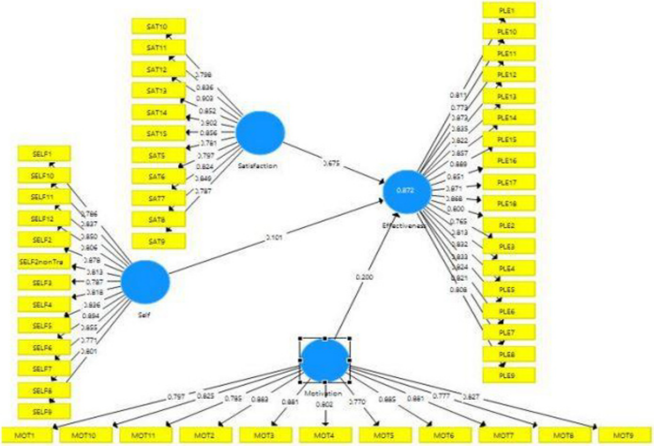


Fig. 5. Measurement Model

Figure 6 shows the Structural Model for Non-mAR Group to show the significant relationship towards Effectiveness.

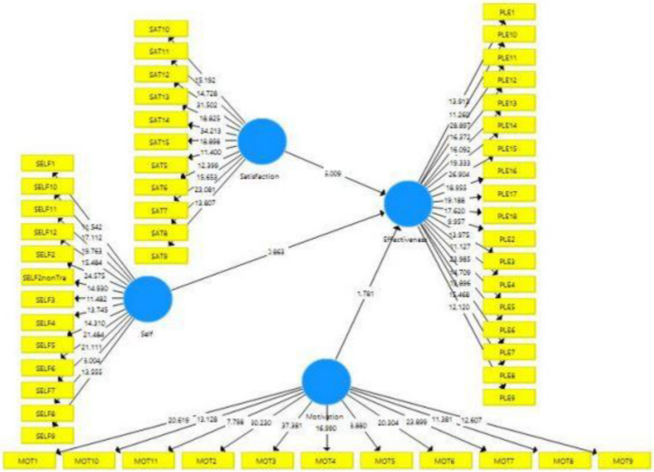


Fig. 6. Structural Model

Table 2 shows the result of the Structural Model, the table shows that there are two significant relationships out of three with Motivation and Satisfaction being the significant relationship with a T-value[a] of 1.781 and 6.009 respectively, and a significant level of 0.038 for Motivation and 0.000 for Satisfaction. Overall, the result shows that the student in the Non-mAR group is Motivated and Satisfied when using the application during 15-minutes self-learning.

Table 2. Significance Testing Result of the Structural Model Path Coefficients

Relationship	**Original Sample (O)**	**Sample Mean (M)**	**Standard Deviation (STDEV)**	**T-Value[a]**	**Significant Level**
Motivation -> Effectiveness	0.200	0.190	0.112	1.781	0.038
Satisfaction -> Effectiveness	0.675	0.663	0.112	6.009	0.01
Self -> Effectiveness	0.101	0.122	0.117	0.863	Not Supported

3.3 mAR Group

In the Mobile-Augmented Reality (mAR) group, the student were given an application installed in their smartphone where they are able to see a virtual 3D model of a Router and a Switch on their table using a Business Card, the business card acts as a Marker to retrieve information to show the 3D model on top of a physical environment. There were 68 students that participated where 56.7% (38 Students) are female while 43.3% (29 Students) were Male. Figure 7 shows the Measurement Model for mAR group where the blue circle shows the Construct of the questionnaire. For mAR group, there is an added construct which is the Features, overall the current relationship of the mAR group are Satisfaction, Self-Efficacy, Motivation and Features towards Effectiveness. The yellow rectangle shows the item that is above 0.6 values. Whereas Fig. 8 shows the Structural Model for mAR Group in order to show the significant relationship towards Effectiveness.

Table 3 shows the Significance testing results for mAR Group. The results show that there are two significant relationships towards Effectiveness, which is Satisfaction and Self-Efficacy with a T-Value[a] of 1.85 and 3.33 respectively. However, the relationship between Features and Effectiveness was very close of being a significant relationship with a T-Value[a] of 1.58, which is close to 1.65. If the features of the mAR application were enhanced, the mAR group could have three significant relationships.

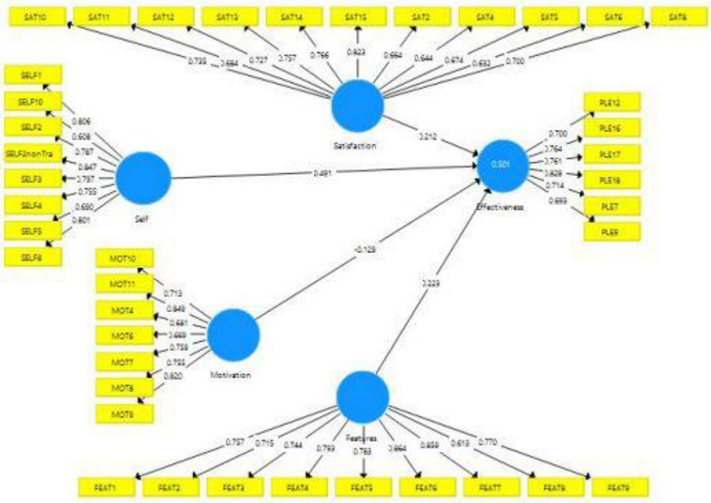


Fig. 7. Measurement Model

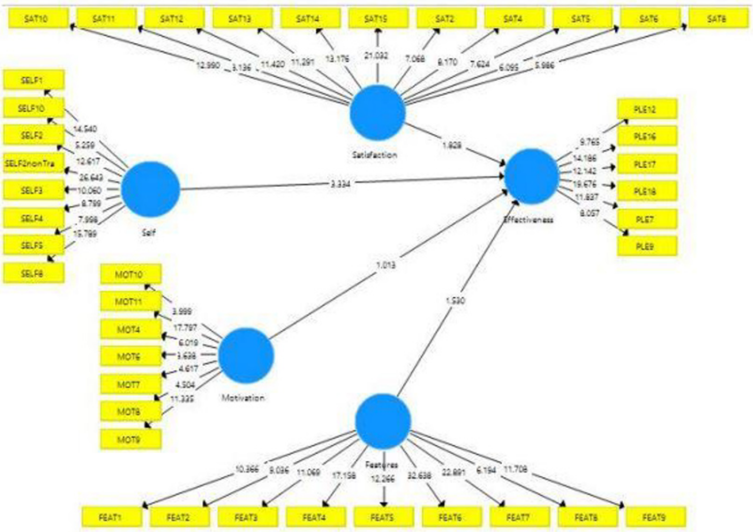


Fig. 8. Structural Model

Table 3. Significance Testing Result of the Structural Model Path Coefficients

Relationship	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T-Value[a]	Significant Level
Motivation -> Effectiveness	-0.140	-0.105	0.142	0.986	Not Supported
Satisfaction -> Effectivness	0.214	0.232	0.116	1.851	0.032
Self -> ffectiveness	0.491	0.489	0.147	3.334	0.000
Features -> Effectiveness	0.227	0.219	0.144	1.580	Not Supported

4 Conclusion

This research shows the result of the three groups that participated in the self-learning environment with three different method and approach towards self-learning, being Traditional, Non-mAR and mAR group. Only Non-mAR and mAR groups has two significant relationships while Traditional group only has one significant relationship, however, mAR group was close to having three significant relationships. With the results produce through the analysis, an enhance framework for mobile-Augmented Reality application was designed to help and guide current developer and researcher to develop and build a better Augmented Reality application [9] in the future to help student learn better with a variety of methods. Figure 10 shows the proposed framework for mAR application [12] as the main contribution resulted from this research, changes on [1] framework include increase learning outcome is resulted in the increase of satisfaction and self-efficacy for a learning environment using mAR. This study mainly focused on student for self-centred learning to ensure that students utilize their smartphone in enriching themselves to learn well.

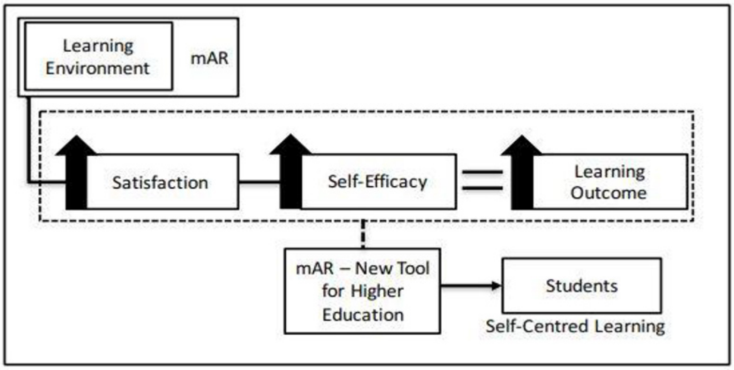


Fig. 9. Proposed Framework for mobile-Augmented Reality Application

References

1. Jamali SS, Shiratuddin MF, and Wong KW. *An overview of mobile-augmented reality in higher education*. International Journal on Recent Trends in Engineering and Technology. 2014 Jul; 11(1):229–38.
2. Ali A, Alrasheedi M, Ouda A & Capretz LF, *A Study of the Interface Usabiility Issues of-Mobile Learning Applications for Smartphones from the User's Perspective.* International Journal on Integrating Technology in Education (IJITE) Vol.3, No.4, December 2014.
3. Abd Majid NA & Husain NK. *Mobile learning application based on augmented reality for science subject: ISAINS*. ARPN Journal of Engineering and Applied Sciences. 2014 Sep; 9(9):1455–60.
4. Wendeson, S., Ahmad, W.F., & Haron, N.S. (2010). Development of Mobile Learning Tool, Information Technology (ITSim). *2010 International Symposium*, 1, 139-144.
5. Barreh KA., & Abas ZW. *A framework for mobile learning for enhancing learning in higher education.* Malaysian Online Journal of Educational Technology. 2015; 3(3):1–9.
6. Hair JF & Sarstedt M & Christian M. Ringle. *An assessment of the use of partial leastsquares structural equation modeling in marketing research.* J. of the Acad. Mark.Sci. (2012) 40:414–433.
7. Lamounier E, Bucioli A, Cardoso A, Andrade A, & Soares A. *On the use of augmented reality techniques in learning and interpretation of caridiologic data*. Annual International Conference of the Institute of Electrical and Electronics Engineers (IEEE) Engineering in Medicine and Biology, Buenos Aires, Argentina; 2010 Aug 31– Sep 4. p. 610–3.
8. Corbeil JR, & Valdes-Corbeil ME. *Are you ready for mobile learning. Educause Quarterly*. 2007; 30(2):51–8.

9. Masrom M. *Implementation of Mobile Learning Apps in Malaysia Higher Education Institutions*. E-Proceeding of the 4th Global Summit on Education 2016, Kuala Lumpur, Malaysia; 2016 Mar. p. 268–76.
10. Aaron FB & Aslina B. Exploring mobile-Augmented Reality in Higher Education. In Proc.4th International Conference on Mathematical Sciences and Computer Engineering (ICMSCE). 4th-5th May 2017, (2017) Langkawi.
11. Aaron FB and Aslina B. A Framework for Developing Mobile-Augmented Reality in Higher Learning Education, Indian Journal of Sciences and Technology (ISSN 0974-5645), 10(39): 1-8, (2017) DOI: 10.17485/ijst/2017/v10i39/119872.
12. Aaron FB. and Aslina B. Mobile-Augmented Reality Framework for Students Self-CenteredLearning in Higher Education Institutions, Journal of Fundamental and Applied Sciences (ISSN1112-9867),10(5S):258-269,(2018) doi:http://dx.doi.org/10.4314/jfas.v10i5s.23

Computational Optimization Analysis of Feedforward plus Feedback Control Scheme for Boiler System

I.M.Chew [1], F.Wong [2], A.Bono [2], J.Nandong [1], and K.I.Wong [1]

[1] Curtin University Malaysia, Sarawak, Malaysia
[2] Universiti Malaysia Sabah, Sabah, Malaysia
chewim@curtin.edu.my

Abstract. Computational optimization via artificial intelligence has been considered as one of the key tools to gain competitiveness in Industrial Revolution 4.0. This paper proposes computational optimization analysis for designing the widely used industrial control systems - feedforward and feedback control schemes. Although several different optimal tunings for servo and regulatory control problems exist, their applications often present some challenges to plant operators. Plant operators often face difficulties to obtain satisfactory PID controller settings by using the conventional tuning methods, which rely heavily on engineering experience and skills. In the proposed intelligent tuning method for the feedforward plus feedback control system, the closed-loop stability region was first established, which then shall provide the upper and lower limits for computational optimization analysis via Genetic Algorithm. Based on a jacketed reactor case study, the performance of feedforward plus feedback control scheme tuned via Genetic Algorithm was compared to that tuned via Ziegler-Nichols tuning. Comparison of performances showed that computational optimization method via Genetic Algorithm gave improved performances in terms of servo and regulatory control objectives.

Keywords: Feedforward plus Feedback, Stability Margin, Genetic Algorithm, Optimal Tuning.

1 Introduction

1.1 Temperature Control of Boiler System

Temperature control is one of the primary control objectives for boiler system operations. Although a single loop feedback system with PID controller is often capable to control the boiler temperature, the closed-loop performance against disturbances can be very poor, i.e., sluggish regulatory performance. The speed of control action is based on the selected control scheme (feedback only or feedforward plus feedback) as well as the applied tuning methodologies. Among the feedback tuning methodologies, the Ziegler-Nichols (ZN) tuning often yields aggressive performance, which sometimes leads to oscillations when the control system is subject to

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_10

setpoint changes and disturbances [1]. In contrast, the IMC-based tuning often results in slower closed-loop response than that of the ZN tuning. Hence, the IMC-based tuning normally does not produce oscillatory closed-loop responses [2]. When only a feedback control structure is used, the corrective action can only begin after the measured process variable has been forced away from the setpoint [2]. Thus, the feedback control scheme may give very poor regulatory control performance even when its servo control performance is satisfactory. In process industry, both feedback and feedforward control schemes are combined in order to obtain fast disturbance rejection and setpoint tracking responses. However, the design or tuning task for the combined feedforward and feedback control scheme is often more challenging than the tuning of a feedback control scheme only. The feed forward control strategy can be implemented by using an additional sensor to measure disturbances of interest [1]. A disturbance model can be constructed and then integrated with the process model to provide an 'early warning' to the involved control system.

1.2 Feedforward plus Feedback Control Scheme

Embedding feedforward into feedback control scheme gives better performance as the effect of a disturbance that can be measured before it affects the process variable. It is widely recommended to the boiler system, which has to deal with great disturbances [2, 3]. Fig. 1 illustrates the block diagram of feedforward plus feedback control scheme.

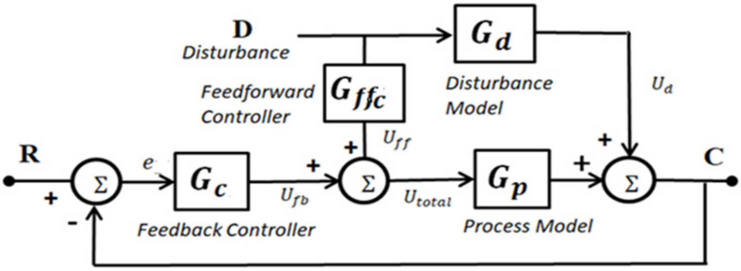


Fig. 1. Block diagram of feedforward plus feedback control system

Feedforward and feedback control scheme are complementary. i.e that each can overcome the disadvantages of the other so that together they are superior to either method alone [4].

In general, feedforward plus feedback control scheme is often used to improve the regulatory control performance of a closed-loop system that is also controlled by a conventional feedback control loop. To obtain a reliable analysis, it is crucial to have accurate models of the process capturing the dynamics of manipulated and disturbance variables in a given process.

For optimal regulatory and servo control performances, two different PID tunings are often required for control objectives. An optimal tuning for the regulatory control objective could lead to a poor servo control performance and vice versa. Because of different control objectives often require different sets of optimal tuning, to find the tuning values for an optimal trade-off between the regulator and servo control objectives can be very challenging as well as confusing to plant operators. The present work attempts to address this challenge in an effective manner through computational optimization analysis approach.

1.3 Introduction to Computational Optimization Analysis with Genetic Algorithm

Computational optimization via artificial intelligence has been considered as one of the key tools to gain competitiveness in Industrial Revolution 4. In this paper, Genetic Algorithm, *GA* as one of the computational optimization analysis is recommended to the feedforward plus feedback control scheme. *GA* is a global searching technique which has been applied to many engineering and optimization problems. It uses a genetic-based mechanism to iteratively generate new solutions from currently available solutions. The powerful capability of *GA* in locating the global optimal solution is used in the design of controllers.

GA was firstly proposed by Holland in 1962 [6, 7]. It is the procedure of adaptive and parallel search for the solution of complex problems. *GA* has been successfully applied to many different problems, such as: traveling salesman [8], graph partitioning problem, filters design, power electronics [9], etc. It has also been applied to machine learning [10], dynamic control system using learning rules and adaptive control [11]. *GA* can be interacted with other Artificial Intelligence techniques, like Fuzzy Sets and Artificial Neural Network, and Multi-Objective Genetic Algorithm and Superheater Steam Temperature Control [12]. In this paper, *GA* has been utilized for finding optimal tuning values of both servo and regulatory control to the feedforward and feedback control loop.

This paper aims to justify the *GA*'s compatibility to obtain optimum PI tunings for both servo and regulatory problems, which is compared to conventional tuning methods in Jacketed Reactor of LOOP-PRO software [5]. The paper is organized as follow: Section 2 develops formulation of feedforward plus feedback control algorithm and safety margin for the stable control. Section 3 explains the working principles of *GA*. Section 4 compares the significant findings for feedforward plus feedback control loop, which is tuned by *GA*, Ziegler-Nichols as well as feedback only control scheme. Last but not least, Section 5 presents the overall conclusion of the studies.

2 Formulation of Feedforward Control System

The mathematical model of the plant (machine, process or organism) can be approximated by using First Order Plus Dead Time, *FOPDT* method. This includes the process as well as disturbance of the plant.

In Fig. 1, feedforward controller $G_{ffc}(s)$ uses the measured value of the disturbance to calculate its feedforward connection, U_{ff}.

The output of Fig. 1 is noted as (1)

$$C = G_d D + G_p U_{total} \tag{1}$$

Develop the expression of U_{total} gives (2)

$$U_{total} = G_{ffc} D + \quad G_c(R - C) \tag{2}$$

Replace (2) into (1) gives (3)

$$C = G_d D + G_p(G_{ffc} D + \quad G_c(R - C)) \tag{3}$$

Simplify the equation yields (4)

$$C = \frac{G_d + G_p G_{ffc}}{1 + G_c G_p} D + \frac{G_c G_p}{1 + G_c G_p} R \tag{4}$$

The stability of the closed-loop system is reflected by the denominator of the transfer function which is known as Characteristic Equation [1, 5]. It is interesting to note that the load changes, feedforward instrumentation, and feedforward controller appears only in the numerator. Thereby, a feedforward algorithm does not destabilize the control, although it potentially leads to poor control due to inability to reduce steady-state offset to zero. The closed-loop transfer function for load changes is explained in (5)

$$\frac{C(s)}{D(s)} = \frac{G_d + G_p G_{ffc}}{1 + G_c G_p} \tag{5}$$

We do expect the "perfect' control where the controlled variable remains exactly at the setpoint despite arbitrary changes in the load variable, D. Thus, the setpoint is constant (R(s) = 0), we want C(s) = 0 even though D ≠ 0, equation above is satisfied as

$$G_d + G_p G_{ffc} = 0$$

Solving G_{ffc} gives the ideal feedforward controller as shown in (6)

$$G_{ffc} = -\frac{G_d}{G_p} \tag{6}$$

G_p and G_d are the process and disturbance model, which can be represented as (7) and (8)

$$G_p = \frac{K_p e^{-\theta_p s}}{\tau_p s + 1} \tag{7}$$

$$G_d = \frac{K_d e^{-\theta_d s}}{\tau_d s + 1} \tag{8}$$

Substituting (7) and (8) into (6), to yield G_{ffc} in (9)

$$G_{ffc} = -\frac{K_d\,(\tau_p s+1)}{K_p\,(\tau_d s+1)}\, e^{-(\theta_d-\theta_p)s} \tag{9}$$

Where,

$Static\ factor,\ G_{ff} = -\frac{Disturbance\ Gain,\ K_d}{Process\ Gain, K_p}$ or known as *Steady State Gain*.

$Lead\ element, \tau_{ld} = \ Process\ Time\ Constant, \tau_p$

$Lag\ element, \tau_{lg} = Disturbance\ Time\ Constant, \tau_d$

For feedforward controller to be realizable, the total deadtime must be a non-negative. In the case of $\theta_d < \ \theta_p$, Erickson [2] suggested to choose θ_d to be similar value with θ_p so to make total deadtime equal to 0.

In tuning feedforward controller, Ogata [15] explained the lead and lag compensation and impact towards disturbance rejection performance. Jobrun [16] recommended to apply feedforward filter improves rejection to disturbance. This paper proposes feedforward ratio, Γ for tuning feedforward algorithm so as to adjust the robustness of feedforward controller as shown in (10)

$$G_{ffc} = -\Gamma\frac{K_d\,(\tau_p s+1)}{K_p\,(\tau_d s+1)}\, e^{-(\theta_d-\theta_p)s} \tag{10}$$

Γ, which is feedforward ratio and $\Gamma \in (0, 1\ )$ a positive scalar parameter which can be used to tune the steady state gain G_{ff}.

3 Genetic Algorithm

3.1 Genetic Algorithm for feedforward plus feedback control loop

Fig. 2 shows the errors from servo and regulatory response has been accumulated for analysis by using *GA*. Generated iterations from optimization analysis is able to give the best PI controller values with Γ for satisfactory performance. The error is measured through Integral Absolute Error (IAE) criterion.

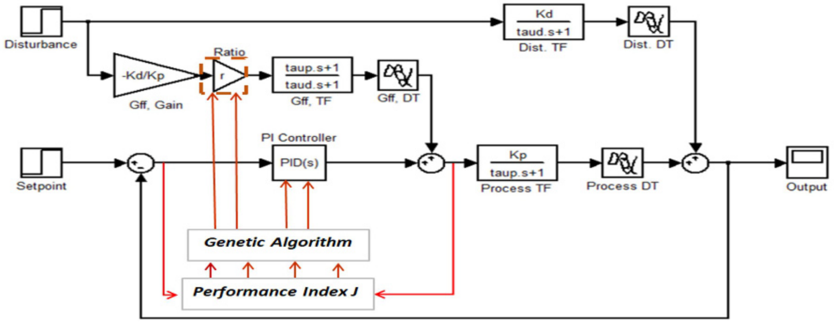


Fig. 2. Block Diagram of computational optimization tuned via *GA*

3.2 Working principle of Genetic Algorithm

GA repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation [13].

The new population contains a large amount of information about the previous generation and carries the new individuals which are superior to the previous generation. It will repeat for many times and the fitness function of all the individuals in the population always increases until certain limit conditions are met. At the end of optimization process, the individual which has the highest degree of fitness are chosen as the optimal solutions of the control terms to be optimized [14]. The flow chart of *GA* is shown in Fig. 3.

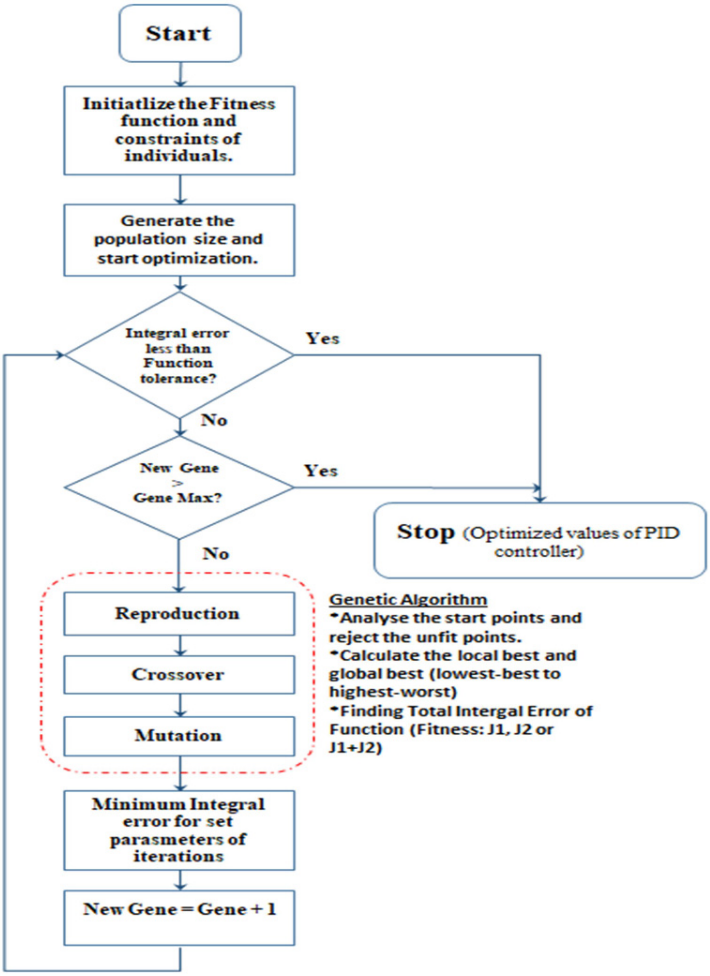

Start
Initiatlize the Fitness function and constraints of individuals.
Generate the population size and start optimization.
Integral error less than Function tolerance?
Yes
No
New Gene > Gene Max?
Yes
Stop (Optimized values of PID controller)
No
Reproduction
Crossover
Mutation
Genetic Algorithm
•Analyse the start points and reject the unfit points.
•Calculate the local best and global best (lowest-best to highest-worst)
•Finding Total Intergal Error of Function (Fitness: J1, J2 or J1+J2)
Minimum Integral error for set parasmeters of iterations
New Gene = Gene + 1

Fig. 3. Flow Chart of Genetic Algorithm

The Scattered function of crossover is selected with ratio of 0.8, which is used to analysis optimum PI tunings whereby mutation is set at Constraint dependent.

4 Analysis and Discussion

4.1 FOPDT models and correlation tunings of PI controller and *Γ*

The *FOPDT* of process and disturbance model are developed through open loop test method, which is obtained from LOOP-PRO software as shown in Table 1.

Table 1. Process and Disturbance Model of Jacketed Reactor (LOOP-PRO)

Process Model	Disturbance Model
$\frac{-0.329e^{-0.721s}}{1.84s+1}$	$\frac{0.8105e^{-1.029s}}{2.268s+1}$

PI controller settings were applied to regulate outlet temperature of the Jacketed Reactor in LOOP-PRO software. The correlation tuning values of PI controller and *Γ* are tabulated in Table 2.

Table 2. PI controller and *Γ* settings of different tuning methods

Tuning Methodology	Proportional Gain, k_c	Integral Time Constant, τ_i	Feedforward ratio, *Γ*
Single loop Control w/o feedforward controller	-0.861	1.84	0
Feedforward plus feedback (IMC)	-0.861	1.84	1
Feedforward plus feedback (Ziegler-Nichols)	-7	2.4	1
Feedforward plus feedback (GA)	-1.564	1.6	0.9

4.2 Feedforward plus feedback control scheme with different feedforward ratio, *Γ*.

It is interesting to note that feedforward plus feedback control scheme significantly improves the regulatory control as shown in Fig. 4. Besides, the response performance can be varied through adjusting the *Γ*. The changes of *Γ* value from 0.2 to 0.8 have steady reduced overshoots in regulatory control however do not effect much to transient response.

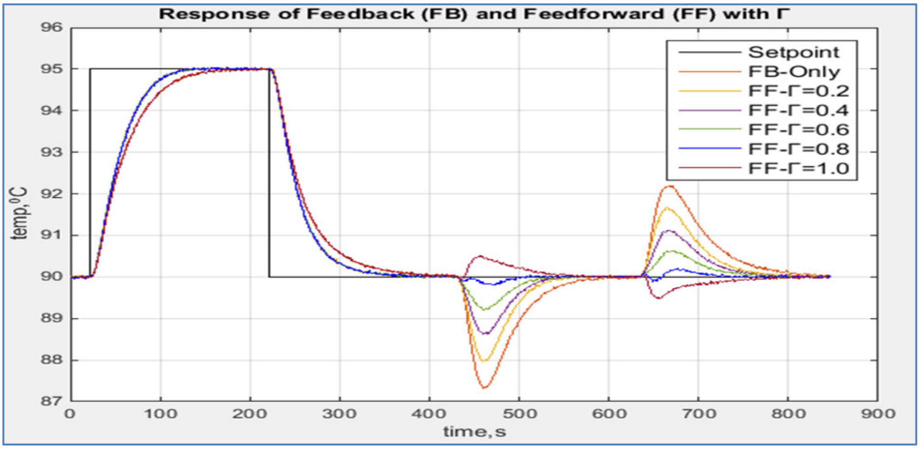


Fig. 4. Feedforward control tuning with varies *Γ* values

4.3 Performance of Computational Optimization Analysis (GA) versus to Conventional Tunings Methodologies

Fig. 5 illustrates relative performance of different tuning methodology in the closed-loop Jacketed-Reactor system. Feedback-only control scheme provides the slowest response in servo control and the largest overshoots in regulatory control. Feedforward plus feedback control scheme significantly improves ability of the PI controller to reject the disturbances. However, it still provides similar transient response for servo control as compared to feedback control scheme.

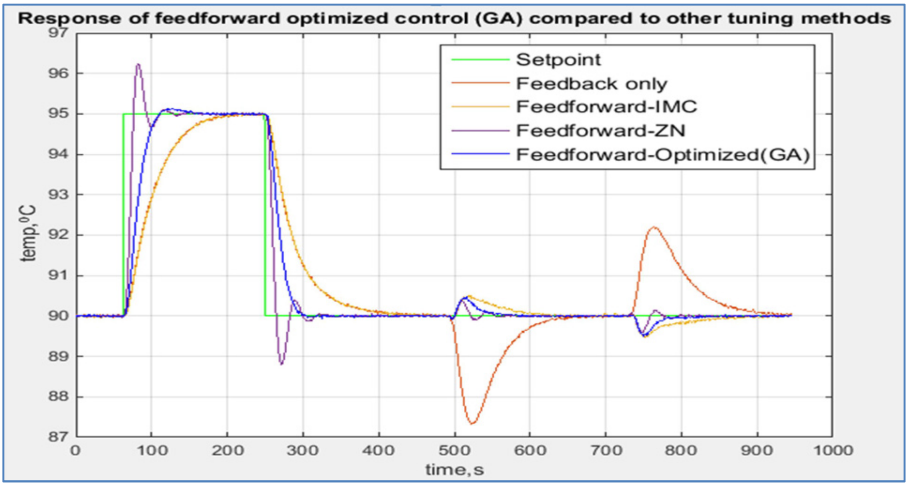


Fig. 5. Performance of varies tuning methodologies

Feedforward plus feedback control scheme tuned by Ziegler-Nichols method has generates the most aggressive response for both servo and regulatory control. However, the result somehow has perturbation and oscillatory therefore less preferably chosen in volatile process such as boiler. In contrast, *GA* provides smoothen responses with minimum oscillations and settling time means higher stability of control.

Table 3 shows the performance indicator of different tunings in feedforward plus feedback control scheme as well as feedback only control scheme. In servo control problems, feedforward plus feedback control tuned by Ziegler-Nichols results the smallest rise time as compared to other tuning methodologies. However, it produces larger overshoots due to high aggressiveness that is potentially destabilizing the control in high volatile boiler system. Feedforward plus feedback control scheme tuned by *GA* had only produces 0.13 ^{0}C of overshoots as well as have the smallest settling time. In regulatory control problems, the overshoots and settling time for *GA* and Ziegler-Nichols are respectively 22s and 24s.

Table 3. Performance Indicator of Servo and Regulatory control.

Tuning Methodology	**Servo Control**			**Regulatory Control**	
	Rise time, s	Overshoot, ^{0}C	Settling time, s	Overshoot, ^{0}C	Settling time, s
Single loop Control w/o feedforward controller	73	0	82	2.63	94
Feedforward plus feedback (IMC)	69	0	80	0.5	35
Feedforward plus feedback (Ziegler-Nichols)	8	1.2	40	0.38	22
Feedforward plus feedback (GA)	27	0.13	34	0.40	24

It is clear that feedforward plus feedback control scheme tuned by GA is preferably selected for safety tuning without worries of controller's setting for fulfilling respective control objective.

5 Conclusion

As found, conventional PI tunings of feedforward plus feedback control scheme improved the regulatory response. The response to disturbance varies based on the setting of Γ. However, the transient responses were not improved.

Feedforward control tuned by Ziegler-Nichols produced the most aggressive response for both servo and regulatory control problem. However, this tuning method is more oscillatory, which led to instability and volatile of boiler system. In contrast, feedforward control tuned by *GA* gives smoothen process responses (smaller oscillations) and did not deteriorated settling time.

It is inspired to conclude that computational optimization technique by GA is able to provide the best PI controller and Γ settings in the prospect of balancing the performance for both servo and regulatory control. The acquired PI settings and Γ setting is k_c= 1.564 , τ_i= 1.6 and Γ = 0.9 for optimal control of feedforward plus feedback control scheme.

References

1. T.E.Marlin, F.: Process Control Designing Process and Control Systems for Dynamic Performance. McGraw Hill Inc, USA(1995).
2. K.T. Erickson, F., J.L.Hedrick, S.: Plantwise Process Control. John Wiley and Sons Inc, USA(1999).
3. R.Kumar, F., S.K.Kingla, S., V.Chopra, T.: Comparison among some well-known control schemes with different tuning methods. Journal of Applied Research and Technology 13, 409-415(2015).
4. J.M.Smith, F., H.C. Van Ness, S., M.M.Aboott, T.: Introduction to Chemical Engineering Thermodynamics. 7th edn. McGraw-Hill, USA(2005).
5. D.Cooper, F.: Practical process control using LOOP_PRO software. Control Station, Inc. USA(2006).
6. J. H. Holland, F.: Outline for a logical theory of adaptive systems. J. ACM, vol. 3, 297-314(1962).
7. J. H. Holland, F.: Adaptation in Natural and Artificial Systems. MI: Univ. Mich. Press, Ann. Arbor(1975).
8. D. E. Goldberg, F., R. Lingle Jr. S. :Alleles, loci and treveling salesman problem," Proc. Int. Conf. Genetic Algorithms and Their Appl.(1985).
9. B.Ozpineci, F., J.O.P.Pinto, S., L.M.Tolbert, T.: Pulse-width optimization in a pulse density modulated high frequency ac-ac converter using genetic algorithms. Proc. of IEEE System, Man and Cybernetics Conf. vol(3), (2001).
10. J. H. Holland, F.: Genetic algorithms and classifier systems: foundations and future directions, genetic algorithms and their applications. Proc. of Sec. Int. Conf. on Genetic Algorithms (1987).
11. P.J.Van Rensburg, F., I. S. Shaw, S., J.D.Van Wyk, T.: Adaptive PID control using a genetic algorithm. Proc. KES'98, Secon. Inter. Conf. Knoledge-Based Intel. Electro. Sys. vol. 2, pp.133-138(1998).
12. A.Y.Begum, F., G.V.Marutheeswar, S. : Genetic Algorithm based Tuning of Controller for Superheater Steam Temperature Control. International Journal of Control Theory and Applications", Vol.10, pp57-65(2017).
13. J. Mchall, F.: Genetic Algorithm for modelling and optimization. Journal of Computational and Applied Mathematics 184, 205-222(2005).
14. R.Malhorta, F., N.Singh, S., Y.Singh, T.: Genetic Algorithms: concepts, design for optimization of process controllers. Canadian Center of Science and Education, Vol.4(2), pp39-54(2005).
15. K.Ogata, F.: Modern Control Engineering. 5th edn. Pearson, USA(2010).
16. J.Nandong, F.: A unified design for feedback –feedforward control system to improve regulatory control performance. International Journal of Control Automation and Systems, 1-8(2015).

Smart Verification Algorithm for IoT Applications using QR Tag

Abbas M. Al-Ghaili[1], Hairoladenan Kasim[2], Fiza Abdul Rahim[2], Zul-Azri Ibrahim[2], Marini Othman[1,2], and Zainuddin Hassan[2]

[1] Institute of Informatics and Computing in Energy (IICE), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia
[2] College of Computer Science and Information Technology (CSIT), UNITEN, 43000 Kajang, Selangor, Malaysia
{abbas, hairol,fiza,zulazri,marini,zainuddin}@uniten.edu.my

Abstract. A Smart Verification Algorithm (SVA) used with Internet of Things (IoT) applications is proposed performing a verification procedure to enable authorized requests by user to access a smart system with help of Quick Response (QR) tag. It uses encrypted QR-tag values to compare them to original values. Three-layers have been proposed for this verification procedure to attain security objectives. The first layer implements a comparison to preserve the system integrated. In the second layer, original values are stored in offline database storage to disable any access caused by threats; to preserve it available. The third one frequently generates an authenticated QR tag using 1-session private key to prevent both information leakage and an unauthorized access if the key was deduced; to keep it confidential. The SVA aims to increase the system privacy. It is evaluated in terms of security factors. Results confirm that it is faster than other competitive techniques. Additionally, results have discussed SVA's robustness against unauthorized access's attempts and brute force attack.

Keywords: Internet of Things, QR Tag, Smart applications.

1 Introduction

Nowadays, researches focusing on the use of QR tag increase rapidly when comparing the current decade to last three decades. The number of recent studies has shown a rapid increment in research studies dedicated for the QR tag related topics especially beyond the year 2010, as shown in Fig. 1. QR tag features (1, 2), such as storing a huge number of encrypted information in a small shaped-image, is one of the reasons attracting many researchers to propose secure systems (3). Some examples include Internet of Things (IoT) based devices (4) intelligent systems (5), smart access cards (6), and automation processes (7) embedded systems (8) that produce smart services to the user and keep data securely stored.

Smart home applications include several purposes raised from the need of use. For example, monitoring application (9), biometrics-based home access system (10)

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_11

…etc. This variety indicates that IoT relies on smart essential tools e.g., QR-tag used as an access tool for IoT applications in which a strong encryption scheme is needed.

QR tag stores a huge number of data in a simple image with small size; many smart systems have exploited such feature(s) so that QR tags are easily scanned. Once, the data stored inside the QR tag has been extracted, the verification process is implemented. The verification process is a very important step to make sure that extracted information is identical to the original one. Thus, the verification process of QR tag is an essential step in smart home applications, because it affects the system privacy.

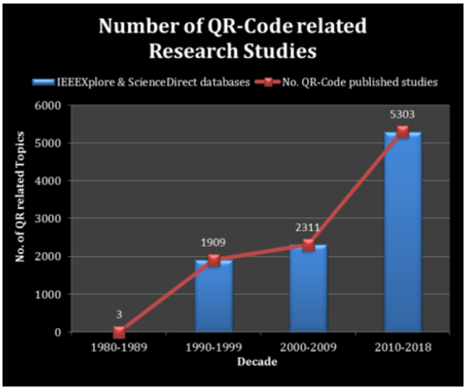


Fig. 1. Number of QR-Code related Research Studies and Topics

However, QR tags are used as different smart tools. The proposed work in (3) has used a QR tag in order to head the robot for measurements. The QR tag acts as a landmark image to ease the recognition process. In (11), the QR tag is generated in such a way to be scanned easily with no much details of black-white-boxes. Other interesting graphical design of QR tag is proposed in (12) to verify documents in terms of authentication and privacy. The original QR tag patterns must be replaced by a specific array of patterns in order for the QR tag to be correctly scanned. In addition, a method to protect the private data stored in the QR tag is designed (13) to overcome the print-and-scan (P&S) operation.

In literature review, there have been many proposed researches using the QR tag with a simple layer of encryption and protection. These methods are suitable for use with smart systems but those which include sensitive data and need a strong protection policy, are vulnerable to threats and attacks. Therefore, to make the QR tag based Smart Verification Algorithm (SVA) achieve a high level of usability and responsiveness, a strong verification policy has been considered in design. In this article, the proposed SVA has considered a number of security issues e.g., integrity and privacy. Additionally, the SVA proposes a simple and reliable QR-tag scanner to verify its contents in terms of authentication. In addition, SVA's verification procedure contains three layers to increase the security. The private key design considers the decryption time caused by an unauthorized action.

This article is organized as follows: Section 2 explains the integrated proposed SVA for IoT applications. Performance analysis and evaluation are discussed in Section 3. Conclusion is provided in Section 4.

2 The Integrated Proposed SVA for IoT Applications

2.1 The Proposed SVA Architecture

The proposed SVA architecture is illustrated in Fig. 2. Once the QR tag is scanned, its information will be verified and then the Reference Value (RV) will be provided to the user to increase the access security. After that, the ID will be checked. Finally, the database is updated based on some security criteria.

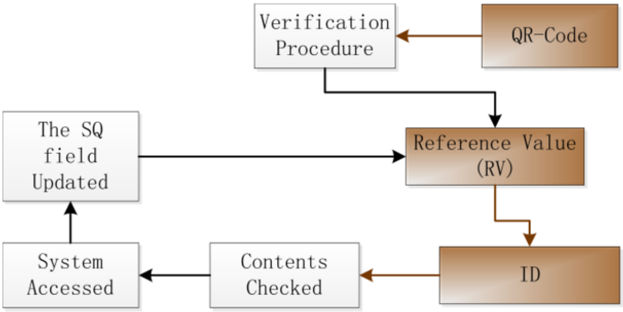


Fig. 2. The Architecture for SVA

2.2 Relation between SVA and Encryption for IoT Applications

After the encryption process has been performed, a hash value is converted to a QR tag logo (as an image) in order to be sent to the user device e.g., mobile app. Then, user is asked to enter an ID value after the system has accepted the QR-tag. Another example of SVA security is that, the RV must be correctly inserted before the access is allowed.

As discussed above, there will be three security levels (QR Scan, RV, ID comparison). Thus, the framework of SVA verification procedure and its relation to the encryption part is illustrated in Fig. 3. The proposed framework has three levels of safety against an unauthorized access and/or modification. In case the system has been modified in an unwanted manner, a wrong RV will be obvious when a hash function is tested. Hence, the system will not allow any access.

This figure shows the SVA relation to the encryption procedure. Once, a user needs to access the system, the encryption procedure immediately is called. The first security step is to show *QR tag by* using a smart device such as a mobile app to allow the user access the system. The pre-generated QR tag which has been stored in the offline database will be used to verify the QR tag with a reference code. If they are identical, the first step is approved; otherwise, the system rejects the process. The second step is to insert an RV to make sure that the right user has displayed the correct QR tag (in the previous step). To increase the security, in this step, the RV is inserted manually by using the mobile app. The RV is frequently changed and the *offline database storage* is updated accordingly. The third security process is *User-id Verification*; it asks the user to enter the *id* number. This number has been previously generated using a complex process. When it is entered, the system will use a special process using a Hash function between the entered and stored ones.

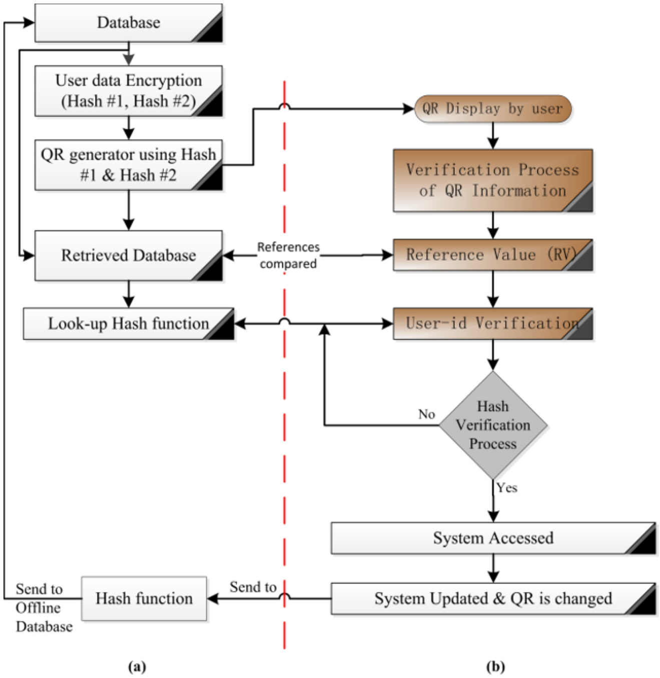


Fig. 3. A SVA Verification Procedure (a) and Encryption (b) Framework

2.3 SVA Flowchart

As discussed earlier, there are four types of verifications which are as follows: a recent face image capturing, fingerprint detection, RV, and user ID. They are discussed in detail as follows:

Face image and Fingerprint Verification Procedure. The first step in SVA is to read information from the offline database and look-up table, once the QR tag has been scanned. The purpose is to do a successful and trust comparison between QR tag and original pre-stored information. The comparison could be illustrated as in Fig. 4. Once information asked by the user and information stored in the database are identical, the system can be accessed. Otherwise, the access is denied.

Hash based Reference Value (HRV) and ID Verifications. The RV is frequently and periodically changed and is expired once it has been used for a short period of time. The SVA needs to update RV in the database. Then, the encryption for the RV will produce a different hash value. The *new* generated QR tag will be updated accordingly. Thus, the new hash value for the user's RV will be periodically verified.

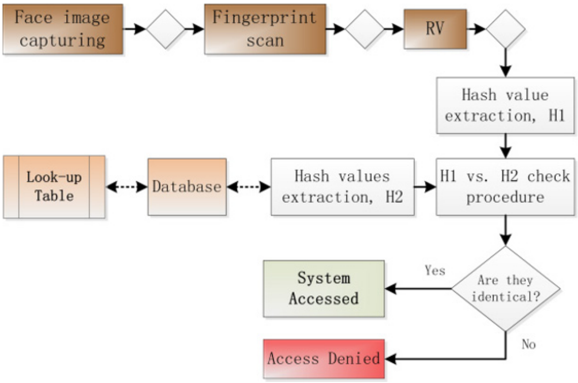


Fig. 4. The Proposed SVA Flowchart

The HRV and ID verifications are applied by using a number of cryptographic operations using algebraic formulas and logic operations. The mathematical procedure is briefly provided in Algorithm 1, Algorithm 2, and a flowchart depicted in Fig. 5.

HRV firstly produces a hash value, h'_{RV}; as expressed in (1):

$$h'_{RV} = hash(RV1 \circledcirc RV2, k_p, M) \qquad (1)$$

whereas

- h'_{RV} represents the first calculated hash value
- $RV1 \circledcirc RV2$ is the hash function input and is a produced mathematically value
- $\circledcirc$ represents the mathematical operations applied on $RV1$ and $RV2$
- k_p is the 2nd input of hash function and is a private key used only one time.
- M is the 3rd input of the hash function and is the message obtained from the user's pre-entered data

Once the value h'_{RV} has been calculated, a further encryption process is applied with a rolling function to produce a new hash value, H_{RV}, as expressed in (2):

$$H_{RV} = E(h'_{RV}, k_s) \qquad (2)$$

whereas

- H_{RV} is the 2nd hash value and is the encrypted value for $RV1$ and $RV2$
- $E(h'_{RV}, k_s)$ encryption procedure applied on h'_{RV} by using a different secret key, k_s.

The proposed algorithm for this procedure is shown in Algorithm 1:

- Start
- Apply pseudorandom number generator (PRNG) on two RVs $RV1$ and $RV2$ values

```
· Apply various mathematical operations to produce $RV1 \circledcirc RV2$ value
· Produce 1-session private key $k_p$
· Implement a hash function ($h'_{RV}$); Eq. (1)
· Implement an encryption algorithm to produce $H'_{RV}$; Eq. (2)
· Apply a rolling function
· End
```

Algorithm 1: HRV Verification Procedures

The pseudo-code of the ID verification procedure is shown in Algorithm 2.

```
· Start
· Do initialization for the following values:
      -Scan QR tag for user #i
      -Ask the user to enter the correct id
      -Call the stored hash value for the entered id(i) → $Hash(id_{Look-up}(i))$
· Do the Hash function for id → $Hash(id(i))$
· Extract the $id_{QR}$ value stored inside the QR tag
      -Do Hash function for this id → $Hash(id_{QR}(i))$
· Compare $Hash(id(i))$ AND $Hash(id_{QR}(i))$ to $Hash(id_{Look-up}(i))$
· Return the comparison value (True OR False)
· End
```

Algorithm 2: ID Verification Procedure

This algorithm checks the user ID collected from different resources. The algorithm compares three collected values. The process compares the encrypted value stored in QR tag, $Hash(id_{QR}(i))$ to the one the user will enter manually which is, $Hash(id(i))$. The result will be compared to the encrypted value stored in the original look-up table, $Hash(id_{Look-up}(i))$.

3 Performance Analysis and Evaluation

This section performs an analysis and evaluation procedure. The performance of the proposed research work is discussed. To achieve a high level of performance of the proposed algorithm, various parameters and factors have been considered. So that points of views are considered to cover most of the security issue.

3.1 Security Factors Analysis

Confidentiality. Information is transmitted between two authorized parties, using a strong encryption algorithm. The QR tag is periodically generated using a 1-session key to increase information confidentiality and keep it secure.

Integrity. QR tag image patterns are scanned and verified. Additionally, face image and fingerprints are verified. If there is any mismatch, the answer will be wrong. Thus, the SVA has no integrity. Thus, a third party has modified the QR tag contents.

Availability. There will be no access by a third party but only one authorized source is allowed to access thru the offline database given a certain period of time.

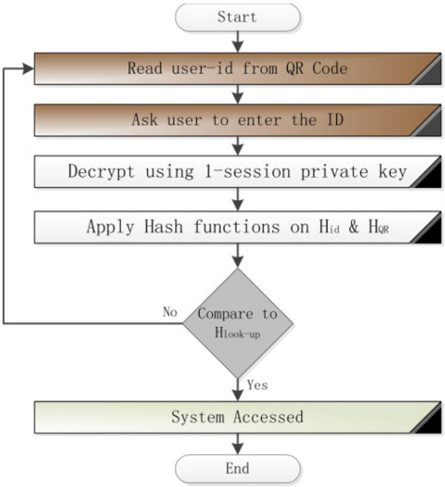

Start
Read user-id from QR Code
Ask user to enter the ID
Decrypt using 1-session private key
Apply Hash functions on H_id & H_QR
Compare to H_look-up
No
Yes
System Accessed
End

Fig. 5. Mathematical operations based user-id comparison flowchart

3.2 Other Factors

Authentication. The authentication factor to be verified is denoted: $F_{authentication}$. Here, a small portion of grey-color patterns (pixels' intensities) are merged with a hash value obtained from the user $Hash(user)$ predefined earlier in a secure mode. The hash value is compared to the original one $Hash(db)$; as mathematically expressed in Eq. (3). If it is correct, a certain hash value will be fit with those grey patterns to produce a new QR tag. The SVA is going to scan the new resulted QR tag. This procedure is performed periodically every 24 hours to guarantee authority of QR tag issue. Any mistake of the entered value will appear in this step. Eq. (8) is applied for QR tag authentication.

$$F_{authentication} = \begin{Bmatrix} 1 & Hash(user) == Hash(db) \\ 0 & Hash(user) \neq Hash(db) \end{Bmatrix} \tag{3}$$

If $F_{authentication} = \{1\}$, then, the QR tag is generated by the original source. Otherwise, it is generated by an unauthorized source. That means there is no authenticity for QR tag being used to access an IoT system.

Robustness. In order for the SVA to allow an access to the related system or physical device by using the application, the entries QR tag Scan, RV, and ID are tested. However, the application might ask the user for information collection procedure. These must be correct; otherwise, the system does not allow any attempt to access. If the QR tag has been used successfully and one or more of other entries was wrong, the user can't access. This verification policy increases the security. Hence, it is clear that any error with SVA input will lead to the same output. Meaning, any possibility of error existence during access's attempts, it is rejected. That can give the SVA more robustness against unauthorized attempts; to prevent threats.

Computation Time based Efficiency. To verify the SVA efficiency, the computation time is calculated and compared to other existing algorithms'.

As discussed earlier on how the encryption process is done to apply a hash function (H_p) on user information. Firstly, using SHA-1, the 1-session key is generated. The hash value extracted from the QR tag (H_{QR}) is compared to the original hash value stored in the offline database; i.e., H_p and H_{QR}. Then, the user is asked to enter an RV and ID value as the message being used for this evaluation. Next, the message, session key, and hash values are encrypted to create a public key. This value is being hashed. Finally, the QR tag will be generated using these information and values. The SVA average computation time compared to other schemes is shown in Table 1.

Table 1. SVA Computation Time Compared to Other Schemes

No. Tests	Password; ms	Certificate; ms	(14); ms	Proposed SVA; ms
10	24.9	45.2	31.4	28.6
20	43.7	91.2	63.1	50.2
50	115.6	212.6	149.0	132.1
100	205.1	452.2	313.3	226.7
200	453.2	921.6	645.7	521.4

This comparison shows that the proposed SVA is faster than Certificate and the work proposed in (14). However, the Password system is faster because the time needed for encryption and verification process of SVA takes more time for a more secure procedure. Another reason is that, the SVA has applied the hash function more than one time; that is to increase the safety of the system being accessed. In Fig. 6, the SVA computation time is on the second rank whereas a less computational time is achieved with a good level of safety and security. In Fig. 7, the average computation time is shown whereas its value ranges between 23.8% and 27.2% among others.

Key Length against brute force attack. The SVA is evaluated in terms of brute force attack

Table 2. Key length based decryption-time evaluation

Key length; size in bits	Number of alternative keys	Time required; 10^6 decryption/µs time in years
168	3.7×10^{50}	6×10^{30}
320	2.1×10^{96}	3.4×10^{76}

Table 2 shows that a very long time is needed to decrypt a message with different key sizes. The last column mentions the time needed for a system in which 10^6 keys could be processed in 1 µs. This is computationally secure.

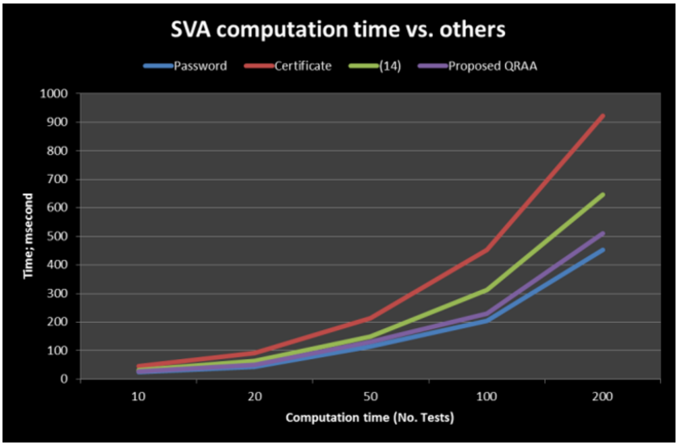


Fig. 6. SVA computation time compared to other methods

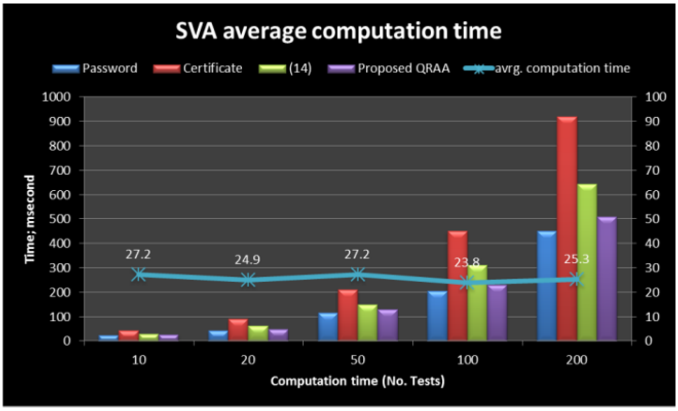


Fig. 7. SVA average computation time

4 Conclusion and Future Works

This paper proposes a simple algorithm used as an access procedure for smart applications and IoT applications such as smart home applications and security gates. The proposed SVA collects user information and encrypts them in order to create a secure QR tag image. The design of SVA has considered the security factors and mechanisms. Results and evaluation have discussed different factors such as integrity, availability…etc. results have confirmed that SVA is strong against brute force attack in terms of time required to crack it. The performance has confirmed it is real-time responsive, reliable, and useable. However, there are some future works suggested. It is recommended, for example, to reduce the computation time during verification process to attain achieving more security and less computation time.

Acknowledgements

Work presented in this paper forms part of the research on Formulation of Evidence Source Extraction Framework for Big Data Digital Forensics Analysis in Advanced Metering Infrastructure, which was partially funded by Universiti Tenaga Nasional Start-Up Grant UNIIG 2016.

References

1. Kirkham T, Armstrong D, Djemame K, Jiang M. Risk driven Smart Home resource management using cloud services. Future Generation Computer Systems. 2014 2014/09/01/;38:13-22.

2. Samuel SSI, editor A review of connectivity challenges in IoT-smart home. 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC); 2016 15-16 March 2016.

3. Nazemzadeh P, Fontanelli D, Macii D, Palopoli L. Indoor Localization of Mobile Robots Through QR Code Detection and Dead Reckoning Data Fusion. IEEE/ASME Transactions on Mechatronics. 2017;22(6):2588-99.

4. Liu Z, Choo KKR, Grossschadl J. Securing Edge Devices in the Post-Quantum Internet of Things Using Lattice-Based Cryptography. IEEE Communications Magazine. 2018;56(2):158-62.

5. Rane S, Dubey A, Parida T, editors. Design of IoT based intelligent parking system using image processing algorithms. 2017 International Conference on Computing Methodologies and Communication (ICCMC); 2017 18-19 July 2017.

6. Huang H-F, Liu S-E, Chen H-F. Designing a new mutual authentication scheme based on nonce and smart cards. Journal of the Chinese Institute of Engineers. 2013 2013/01/01;36(1):98-102.

7. Xiao-Long W, Chun-Fu W, Guo-Dong L, Qing-Xie C, editors. A robot navigation method based on RFID and QR code in the warehouse. 2017 Chinese Automation Congress (CAC); 2017 20-22 Oct. 2017.

8. Ghaffari M, Ghadiri N, Manshaei MH, Lahijani MS. P4QS: A Peer-to-Peer Privacy Preserving Query Service for Location-Based Mobile Applications. IEEE Transactions on Vehicular Technology. 2017;66(10):9458-69.

9. Chen YH, Tsai MJ, Fu LC, Chen CH, Wu CL, Zeng YC, editors. Monitoring Elder's Living Activity Using Ambient and Body Sensor Network in Smart Home. 2015 IEEE International Conference on Systems, Man, and Cybernetics; 2015 9-12 Oct. 2015.

10. Kanaris L, Kokkinis A, Fortino G, Liotta A, Stavrou S. Sample Size Determination Algorithm for fingerprint-based indoor localization systems. Computer Networks. 2016 2016/06/04/;101:169-77.

11. Lin SS, Hu MC, Lee CH, Lee TY. Efficient QR Code Beautification With High Quality Visual Content. IEEE Transactions on Multimedia. 2015;17(9):1515-24.

12. Tkachenko I, Puech W, Destruel C, Strauss O, Gaudin JM, Guichard C. Two-Level QR Code for Private Message Sharing and Document Authentication. IEEE Transactions on Information Forensics and Security. 2016;11(3):571-83.

13. Lin PY. Distributed Secret Sharing Approach With Cheater Prevention Based on QR Code. IEEE Transactions on Industrial Informatics. 2016;12(1):384-92.

14. Kim YG, Jun MS, editors. A design of user authentication system using QR code identifying method. 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT); 2011 Nov. 29 2011-Dec. 1 2011.

Daily Activities Classification on Human Motion Primitives Detection Dataset

Zi Hau Chin[1], Hu Ng[1](✉), Timothy Tzen Vun Yap[1] , Hau Lee Tong[1],
Chiung Ching Ho[1] and Vik Tor Goh[2]

[1] Faculty of Computing & Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
[2]Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia
zihau27@gmail.com, {nghu, timothy, hltong, ccho, vtgoh}@mmu.edu.my

Abstract. The study is to classify human motion data captured by a wrist worn accelerometer. The classification is based on the various daily activities of a normal person. The dataset is obtained from Human Motion Primitives Detection [1]. There is a total of 839 trials from 14 activities performed by 16 volunteers (11 males and 5 females) ages between 19 to 91 years. A wrist worn tri-axial accelerometer was used to accrue the acceleration data of X, Y and Z axis during each trial. For feature extraction, nine statistical parameters together with the energy spectral density and the correlation between the accelerometer readings are employed to extract 63 features from the raw acceleration data. Particle Swarm Organization, Tabu Search and Ranker are applied to rank and select the positive roles for the later classification process. Classification is implemented using Support Vector Machine, *k*-Nearest Neighbors and Random Forest. From the experimental results, the proposed model achieved the highest correct classification rate of 91.5% from Support Vector Machine with radial basis function kernel.

Keywords: Daily Activities, Accelerometer, Classification.

1 Introduction

The recognition of daily activities has received wide attention over the last few decades [2]. With accurate recognition of human activity, devices or robots are able to support humans in an ambient intelligence environment. Besides that, activity recognition system is able to support the unwell and disabled, for instance, observing patient actions in home-based rehabilitation.

Typically, activity recognition system is separated into two approaches, vision-based and sensor based. Vision-based approach utilizes various cameras to capture the agents' activity. Sensor-based approach utilizes sensors that are attached to the body of the agent to gather data of the agents' behavior alongside with the environment [2].

In this research work, we use of the motion data from a wrist worn tri-axial accelerometer, obtained from Human Motion Primitives Detection Dataset (HMP) [1]. For

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_12

feature extraction, nine statistical parameters together with the energy spectral density and the correlation between the accelerometer readings are applied to extract 63 features from the raw acceleration data. Feature selection is executed by applying Particle Swarm Organization (PSO), Tabu Search (Tabu) and Ranker whereas classification is implemented using Support Vector Machine (SVM), *k*-Nearest Neighbors (*k*-NN) and Random Forest.

In this research work, the related works on activity recognition system is presented in Section 2. In Section 3, the proposed model is introduced. Experimental setups are discussed in Section 4. The classification outcomes are revealed in Section 5 together with the analysis to justify the findings. Finally, Section 6 concludes the paper.

2 Related Works

The processes of activity recognition can be broken down to four basic tasks as shown below [2]:

a. Determine the appropriate approach to capture the activities data.
b. Gather, store and process the obtained data by knowledge representation and reasoning.
c. Build computational model and perform analysis using software system.
d. Construct and select the suitable machine learning algorithm to recognize activities.

Action recognition is generally categorized into two different approaches, mainly the vision-based and sensor-based.

In vision-based action recognition system (VARS), still images or videos that are captured by camera are used to recognize actions. This approach utilizes the computer vision to analyse the captured content for action recognition. It plays important role in applications such as surveillance, entertainment and medical. Gaglio et al. [3] applied CMOS colour sensor and CMOS infrared sensor (3D depth sensing) to captured movement of body joints for posture identification. Eum et al. [4] used infrared thermal camera to capture images in dim environment for human detection.

Sensor-based action recognition system (SARS) utilizes various sensors to observe an agent's activity alongside with the changes in the environment [5]. Normally, sensors are attached to human body such as wrist, knee, ankle, chest, waist and head [6]. Sensors such as accelerometers, magnetometers, gyroscopes and vital signs devices are used to obtain angular rates, magnetic field, data of force, temperature and heart beat rate [7]. Ward et al. [7] proposed microphones and tri-axial accelerometers on the wrist and upper arm to obtain the data of sound and motion for assembly and maintenance activities such as drilling, sawing and grinding. Parkka et al. [8] implemented a Global Positioning System by combination of different sensors to capture heart beat and breathe rates.

As SARS is not affected by illumination and occlusion by objects, it is able to function as an alternative to VARS when it fails to perform effectively. Ho et al. [9] implemented both SARS and VARS to build a biometrics system for human recognition. The combination of both SARS and VARS can be a valuable complement to the usage of multimodal biometric systems. However, our research also utilizes SARs method, but with the addition of statistical features. The following Sections detail our methods, results and findings.

3 Proposed Model

This section illustrates the methodology to develop the proposed model. Dataset acquisition, feature extraction, normalization, feature selection and classification methods are expanded in the subsequent sub-sections.

3.1 Human Motion Primitives Detection Dataset (HMP)

Human Motion Primitives Detection Dataset (HMP) is composing of 839 trials of 14 daily activities, performed by 16 volunteers (11 men and 5 women) with ages between 19 to 81 years. The volunteers performed the activities while wearing a wrist-worn tri-axial to his/her right wrist, whereas the supervisor will classify the acceleration data acquired according to the motion. Table 1 shows the activities in HMP.

Table 1. Activities in HMP [1]

Activities of Daily Living	Human Motion Primitives
Toileting	Brush teeth Comb hair
Transferring	Get up from the bed Lie down on the bed Sit down on a chair Stand up from a chair
Feeding	Drink from a glass Eat with fork and knife Eat with spoon Pour water into a glass
Mode of transportation (indoor)	Climb the stairs Descend the stairs Walk
Ability to use telephone	Use the telephone

3.2 Feature Extraction

For feature extraction, nine statistical parameters (time and frequency domain versions – 9 × 3 axes × 2 domains) together with the energy spectral density (3 axes × 2 domains) and correlation between the accelemeter readings (X, Y and Z) are applied to extract 63 (54 + 6 + 3) features from the raw acceleration data. There are 9 statistical parameters extracted from the raw acceleration data are listed in Table 2.

A correlation between X and Y is to measure the strength of the relationship between X and Y. The population correlation coefficient $\rho_{X,Y}$ between two random variables A and B with expected values μ_A and μ_B and standard deviations σ_A and σ_B is defined as

$$\rho_{A,B} = \frac{E[(A-\mu_A)(B-\mu_B)]}{\sigma_A \sigma_B} \tag{1}$$

where E is the expected value operator. The energy of a signal is represented by the strength of the signal. The energy En of a signal $x(t)$ is defined as

$$En = \int_{-\infty}^{\infty} |x(t)|^2 \, dt \tag{2}$$

Table 2. 13 Parameters extracted from acceleration data

Parameters	Description
Minimum	Lowest value in a range of value.
Maximum	Highest value in a range of value.
Standard deviation	Amount of deviation for a group.
Median	Value separating the higher half data sample from the lower half.
Mean	Average of the numbers.
Skewness	Measurement of the lack of symmetry.
Kurtosis	Sharpness of the peak of frequency-distribution curve.
Absolute Skewness	Absolute value of skewness.
Absolute Kurtosis	Absolute value of kurtosis.

3.3 Normalization and Feature Selection

Normalization was implemented on the extracted features to confirm that they are normalized and not biased. The intention is to avoid cases where more weights are given implicitly to features with larger scales than those with smaller scales. Linear scaling is applied to rescale the extracted features to the scope between 0 and 1.

The aim of feature selection is to lower the amount of features by including and excluding the features in the data without imposing changes on it, whereas a dimensionality reduction lowers the number of features by forming a new set of grouping of features. In this paper, PSO, Tabu and Ranker were used to find those features that contributed positive roles in the classification process. There are chosen as they are found to perform effectively in pattern recognition.

In PSO, possible solutions known as particles will flow in the hyperspace. At first, random velocity and position are assigned to each individual particle to initialize the selection. By every iteration, the velocity of each particle will change and accelerate owards the best local and global classification rate [10]. Tabu [11] implements the local search methods by searching for an improved solution by checking at its immediate neighbors. A short-term memory list is employed to help guide and record the process of searching. Ranker chooses the best feature set based on their individual assessment in a ranking. It assesses the weight of an individual feature, where the higher the rank, the more important a feature is. It assesses the worth of an attribute by evaluating the correlation between attribute and the class [12].

3.4 Classification

For classification, three classifiers were used, namely *k*-nearest neighbors (*k*-NN) with Euclidean distance metrics, Random Forest (RF) and Support Vector Machine (SVM).

k-NN is a non-parametric classifier where the neighbors of an object will cast a majority vote to determine the object's class [13]. The object will be labeled to a particular class by referring to the majority poll of nearest neighbors. For this work, the value of neighbor, *k* (number of nearest neighbors) is the only parameter manipulated. RF [14] creates decision tree based on the random selection of data subsets and variable subsets. A voting is formed after each of the observation was conducted. The class with a highest vote will be the class of the object. For this work, three parameters were manipulated, seeds (*S*), number of iterations (*I*) and number of randomly chosen attributes (*A*). In SVM, *n*-dimensional space (*n* = the number of features) is plotted with points which represent data item, then the value of each feature will be represented as a certain coordinate [15]. SVM will find the hyper-plane that will accurately differentiate both the classes. The best hyper-plane will be the one that maximizes the margins from both classes. For this work, three different kernels were used, which is linear (Ln), polynomial (Poly) and radial basis function (RBF). The parameters for Ln kernel is cost (*C*); RBF kernel are cost (*C*) and gamma (*G*); Poly kernel are cost (*C*), gamma (*G*), coefficient (*R*) and degree (*D*).

3.5 Performance Evaluation

Ten fold cross validation was used throughout the training and testing process. By splitting all the features vectors from the dataset into ten distinctive subsets, where nine subsets were used for training while one subset was used for testing. By repeating the iteration 10 times, where all the features vectors of each disjointed subsection are ordered into classes during the validation test. Later, the classification rate is calculated by average out the cross validation outputs. For this work, correct classification rate (CCR) was used to evaluate the percentage of activities that are correctly classified by the classifier.

4 Experiments Set Up

The experiments were conducted in two phases: training and testing. All 839 trials from HMP were fully utilized in the experiments. In this research, three feature selection techniques were implemented; PSO, Tabu and Ranker. For classification, three classifiers were implemented; *k*-NN, RF and SVM. The training was conducted to find the models for classification, the parameter values that provided the best result for each classifier was recorded and act as the models for testing phase, which are revealed in Table 3. In the testing section, the model acquired from the training phases is applied for activities classification for HMP.

Table 3. Parameter values for each classifier

Classifiers	Parameters
k-NN	$k = 3$
RF	A =0, S = 10, I = 100
SVM with Ln kernel	$C = 64$
SVM with Poly kernel	$C = 6144$, $G = 0$, $D = 3$, $R = 0.0$
SVM with RBF kernel	C = 4096, $G = 1.0$

5 Result and Discussion

To assess the accomplishment of the proposed model on the HMP, various experiments were carried out. This section shows and discusses the outcomes of these experiments which were intended to evaluate the classification results of the proposed model. Table 4a and Table 4b summarize the classification result of the classifiers in conjunction with the feature selectors. Table 5 shows the average CCR (%) obtained for each feature selection technique with various classifiers.

Table 4a. Classification Result

	Classifiers with various feature selectors								
		***k*-NN**			**RF**			**SVM Ln kernel**	
	PSO	Tabu	Ranker	PSO	Tabu	Ranker	PSO	Tabu	Ranker
CCR(%)	87.2	89.4	88.9	90.0	90.7	90.6	89.6	88.7	88.6
Average		**88.5**			**90.4**			**89.0**	

Table 4b. Classification Result

	Classifiers with various feature selectors					
		SVM Poly kernel			**SVM RBF kernel**	
	PSO	Tabu	Ranker	PSO	Tabu	Ranker
CCR(%)	89.6	90.5	90.4	91.5	90.8	90.9
Average		**90.2**			**91.1**	

Table 5. Average CCR (%) for each feature selection technique

	PSO	**Tabu**	**Ranker**
k-NN	87.2	89.4	88.9
Random Forest	90.0	90.7	90.6
SVM Ln kernel	89.6	88.7	88.6
SVM Poly kernel	89.8	90.5	90.4
SVM RBF kernel	91.5	90.8	90.9
Average	**89.6**	**90.0**	**89.9**

As shown in Tables 4a and 4b, SVM with RBF kernel yields the highest average CCR (91.1%). As SVMs specialized in finding hyperplane that will accurately differentiate both the classes, by maximizing the margin form both classes. It will ignore all the other data points if optimum hyperplane was found in a linearly separable problem, thus lowering the number of selected support vectors and suitable to solve problem with large number of features [15]. Finding from Hyun and Lee [16] shows SVM RBF generally outperforms Ln kernel and Poly Ln due to its flexibility that allows more functions to be model within its function space.

From Table 5, it can be found that the top performance was observed in Tabu, where only 29 extracted features were selected for classification, with comparing to 32 (PSO) and 63 (Ranker). This shows that Tabu is more effective in selecting features that can contribute positively to the classification process.

Among those selected features by the three feature selectors, the top six ranked features were captured from acceleration data in X and Z axis. Three of them are median, minimum and mean of X axis (vertical movement of the right wrist). These features are prominent in determining activities such as “Sit down” and “Stand up”, which were found to have massive change of acceleration in the X axis. The remaining three features are median, mean and standard deviation of the Z axis (moving the right wrist perpendicularly towards or leaving the body). These features are prominent in determining activities such as “Lie down”, “Sit down” and “Stand up”.

The generated confusion matrix of the classification process with SVM with RBF kernel and PSO (highest CCR) is shown in Table 6. From Table 6, it is observed that activities such as “Descend the stairs”, “Climb the stairs” and “Walk” produced lower recognition with high misclassification. This is due to the three activities having high similarities in the acceleration signals of the right hand, but not distinct enough for movements involving lower parts of the body [8, 17].

Table 6. The generated confusion matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	classified as
11	1	0	0	0	0	0	0	0	0	0	0	0	0	a = Brush teeth
0	91	0	2	0	0	0	1	0	0	0	0	0	8	b = Climb the stairs
0	0	31	0	0	0	0	0	0	0	0	0	0	0	c = Comb hair
0	5	0	36	0	0	0	0	0	0	0	0	0	1	d = Descend the stairs
0	0	0	0	98	0	0	0	0	1	0	1	1	0	e = Drink from a glass
0	0	0	0	0	5	0	0	0	0	0	0	0	0	f = Eat with fork and knife
0	0	0	0	0	0	3	0	0	0	0	0	0	0	g = Eat with spoon
0	0	0	0	0	0	0	93	7	0	1	0	0	0	h = Get up from the bed
0	0	1	0	0	0	0	6	14	0	6	1	0	0	i = Lie down on the bed
0	0	0	0	0	0	0	0	0	100	0	0	0	0	j = Pour water into a glass
0	0	0	0	0	0	0	3	4	0	89	4	0	0	k = Sit down on a chair
0	2	0	0	0	0	0	5	0	1	6	86	0	2	l = Stand up from a chair
0	0	0	0	0	0	0	0	0	0	0	0	13	0	m = Use the telephone
0	6	0	0	0	0	0	1	0	0	2	0	0	91	n = Walk

The size of the training model will also affect the classification result. In our case, since the number of trials for “Lie down on the Bed” was significantly lesser (22 trials) compared to the number of instances for other activities, thus the training sample size of it was low and in turn affected the classification result. Moreover, volunteers do not move their hand much when performing this activity, so the obtained acceleration data was not distinct.

On the other hand, activities such as “Brush teeth”, “Drink from a glass”, “Pour water into a glass”, “Comb hair” are heavily involved the hand movement, primarily where the tri-axial accelerometer was worn, were correctly classified with an average accuracy of 98.53% throughout the experiment. These activities have distinct properties of acceleration data that allowed them to be easily classified.

From our findings, additional sensors are suggested to be worn at different parts of the body for activities involving both hands and the lower body, in order to improve the classification performance.

6 Conclusion

This research performed feature extraction to extract statistical parameters, energy spectral density and correlation between the accelerometer readings from Human Motion Primitives Detection dataset. Feature selection techniques such as PSO, Tabu Search and Ranker and classifier such as *k*-NN, Random Forest and SVM were implemented. A number of experiments have been carried out to evaluate the performance of the proposed solution. The proposed model was proved to be able to correctly classify 14 activities that performed by 16 volunteers.

7 Acknowledgement

The authors would like to thank Bruno et al. from Università degli Studi, Geneva for offering the use of the database in this research. Financial support from Multimedia University under the Multimedia University Capex Fund with Project ID MMUI/CAPEX170008, & the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme with grant number FRGS/1/2015/SG07/MMU/02/1, and TM R&D (UbALive) are gratefully acknowledged.

8 References

1. Bruno, B., Mastrogiovanni, F., & Sgorbissa, A.: A public domain dataset for ADL recognition using wrist-placed accelerometers. In the 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 738-743. IEEE, Scotland (2014).
2. Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z.: Sensor-based activity recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), 790-808 (2012).
3. Gaglio, S., Re, G. L., & Morana, M.: Human activity recognition process using 3-D posture data. IEEE Transactions on Human-Machine Systems, 45(5), 586-597 (2015).
4. Eum, H., Lee, J., Yoon, C., & Park, M.: Human action recognition for night vision using temporal templates with infrared thermal camera. In 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 617-621. IEEE, Korea (2013).

5. Chen, L., & Nugent, C.: Ontology-based activity recognition in intelligent pervasive environments. International Journal of Web Information Systems, 5(4), 410-430 (2009).
6. Long, X., Yin, B., & Aarts, R. M.: Single-accelerometer-based daily physical activity classification. In Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6107-6110. IEEE, Minneapolis (2009).
7. Ward, J. A., Lukowicz, P., Troster, G., & Starner, T. E.: Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Transactions on Pattern Analysis And Machine Intelligence, 28(10), 1553-1567 (2006).
8. Parkka, J., Ermes, M., Korpipaa, P., Mantyjarvi, J., Peltola, J., & Korhonen, I.: Activity classification using realistic data from wearable sensors. IEEE Transactions on information technology in biomedicine, 10(1), 119-128 (2006).
9. Ho, C. C., Ng, H., Tan, W. H., Ng, K. W., Tong, H. L., Yap, T. T. V., Chong, P.F., Eswaran, C. & Abdullah, J.: MMU GASPFA: a COTS multimodal biometric database. Pattern Recognition Letters, 34(15), 2043-2050 (2013).
10. Kennedy, J.: Particle swarm optimization. In Encyclopedia of machine learning, pp. 760-766. Springer US. Kennedy (2011).
11. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers & operations research, 13(5), 533-549 (1986).
12. Hall, M. A., & Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering, 15(6), 1437-1447 (2003).
13. Altman, N. S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185 (1992).
14. Ho, T. K. : Random decision forests. In Proceedings of The Third International Conference on Document Analysis and Recognition (vol. 1, pp. 278-282). IEEE, Montreal (1995).
15. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E.: Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190 (2007).
16. Byun, H., & Lee, S. W.: Applications of support vector machines for pattern recognition: A survey. In Pattern Recognition With Support Vector Machines, pp. 213-236. Springer, Berlin, Heidelberg (2002).
17. Chernbumroong, S., Atkins, A. S., & Yu, H.: Activity classification using a single wrist-worn accelerometer. In 5th International Conference on Software, Knowledge Information, Industrial Management and Applications (SKIMA), pp. 1-6. IEEE, Benevento (2011).

Implementation of Quarter-Sweep Approach in Poisson Image Blending Problem

Jeng Hong Eng[1], Azali Saudi[2] and Jumat Sulaiman[3]

[1,3] Faculty of Science and Natural Resources, Universiti Malaysia Sabah,
88400 Kota Kinabalu, Sabah, Malaysia
[2] Knowledge Technology Research Unit, Faculty of Computing and Informatics,
Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
jenghong93@gmail.com, azali@ums.edu.my and jumat@ums.edu.my

Abstract. The quarter-sweep scheme has been used in solving boundary value problems efficiently. In this paper, we aim to determine the capability of the family of Gauss-Seidel iterative methods to solve the Poisson image blending problem, which are the Full-Sweep Gauss-Seidel (FSGS), Half-Sweep Gauss-Seidel (HSGS) and Quarter-Sweep Gauss-Seidel (QSGS). Second order finite difference approximation is used for the discretization of Poisson equation. Finally, the numerical results show that QSGS iterative scheme is more competent as compared with the full- and half-sweep approaches while obtaining the same quality of output images.

Keywords: Quarter-sweep iteration, Poisson image blending, Poisson equation.

1 Background

Poisson image blending is one of the fundamental classes of problem in image processing. It named as Poisson image blending because it involves the solving of Poisson equation in the process. Poisson image blending is used to create a new desired image from a set of source and destination images. This concept of image blending process by solving Poisson equation was initiated by [1] and it is based on the gradient approach instead of the pixels.

The development of this concept had inspired some researchers to improve some issues caused from the blending process by using different ways, for example the issue of execution time and color inconsistency. Therefore, author in [2] proposed to add another boundary condition in the selected region. With this two boundary conditions in the selected region, the issue of color inconsistency in the output images is resolved.

In addition, Fourier method is suggested to shorten the execution time and ease the implementation process by [3]. Fourier method is a type of non-iterative method. In their paper, unconstrained boundary condition is used for the Poisson equation and the desired region is automatically chosen by an algorithm.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_13

The classic gradient domain method which is only considered the boundary pixels of destination image resulted in bleeding artifacts issue. Thus, [4] recommended modified Poisson blending method to solve this problem by examining the boundary pixels of both source and destination images in the blending process. Then, the operation of alpha compositing is implemented at the final step.

Recent research by [5] and [6] employed the subdivide image and generative adversarial networks (GAN) approaches in Poisson image blending. In [5], the researchers subdivide the desired region into small pieces before the compositing process and this had reduced the cost of computational time. Meanwhile, researchers in [6] presented a new concept by utilizing the Gaussian-Poisson equation and GAN for image blending. Their method successfully generates a high-resolution and realistic image.

Besides in image editing, image processing also has it wide applications in medical imaging, for example in detecting the cancer cells in body [7] and enhancing the medical X-Ray images [8]. Furthermore, it applied in agriculture to classify the types, sizes and colors of the crops [9, 10].

In this paper, we focus on solving Poisson image blending problem by using numerical approach. The concept of quarter-sweep in solving boundary value problem was proposed by [11]. It is then been applied in solving elliptic equations by [12, 13]. However, quarter-sweep concept has not been applied in any Poisson image blending problem. There are some researches applied full and half-sweep concepts in Poisson image blending problem, for example in [14-16].

Thus, we aim to determine the efficiency of quarter-sweep approach in solving proposed problem. The numerical results obtained are compared with the results obtained by using full and half-sweep approaches. In addition, the parameters used to measure the efficiency of the proposed method are the number of iterations and the compositing time. On the other hand, the robot path planning problem is also employing the Laplace equation which can be derived from Poisson equation [17, 18].

2 Poisson Image Blending

Two digital images are involved in the process of image blending, the source and destination images. Fig. 1 illustrates the coordinate system of a digital image.

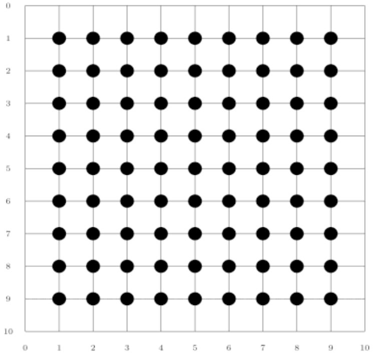


Fig. 1 Finite grid network of a digital image in computer screen.

The pixels are stored inside the image as a two dimensional array, represented by •. In this paper, RGB color model is used. Thus, the proposed problem is solved three times for the three color channels and merged to generate the final images.

According to [1], Poisson image editing is using the idea of interpolation with the guidance vector field selected by the user and its solution is defined as a minimization problem, as follows,

$$min_q \iint_{\mathrm{B}} |\nabla q - \mathbf{f}|^2 \, with \; q|_{\partial \mathrm{B}} = q^*|_{\partial \mathrm{B}} \qquad (1)$$

where B is the selected region with its boundary $\partial \mathrm{B}$ from the source image s while q and q^* are the output and target images respectively. In addition, **f** is a vector field. The author in [1] stated that **f** is the gradient of some function when it is conservative. The first step in Poisson image blending process is to select the desired region from the source image. Then the desired region is cloned into the target image to generate the new output image.

The solution of the minimization problem (1) is a new set of intensity values that minimized the difference between the vector field and the gradient of the new image. To obtain the new set of intensity values, Poisson equation with Dirichlet boundary condition is solved because their solutions are equivalent,

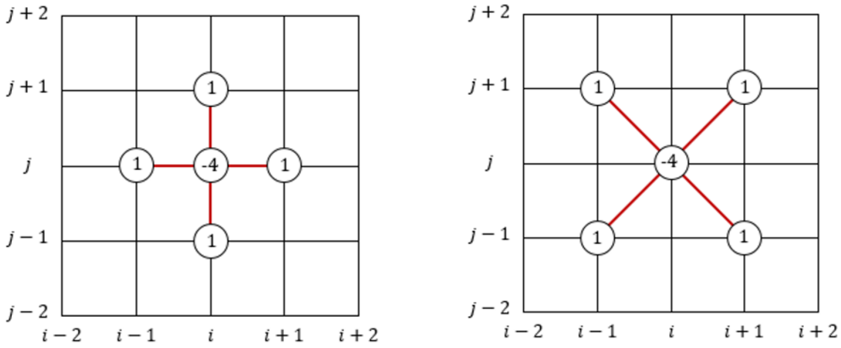
$$\Delta q = \Delta s \; at \; \mathrm{B} \; with \; q|_{\partial \mathrm{B}} = q^*|_{\partial \mathrm{B}} \qquad (2)$$

The vector field is directly generated from the source image and Δ is the Laplacian operator.

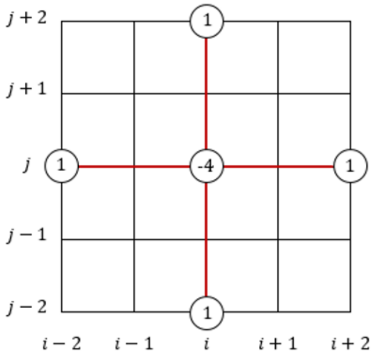
Finite difference method is the most suitable numerical method to discretize the Poisson equation because it is an elliptic partial differential equation with regular domain. Three types of five-point Laplacian operator are used which are the full-sweep operator with grid spacing h, half-sweep operator with grid spacing $\sqrt{2}h$ and quarter-sweep operator with grid spacing $2h$. All these operators are applied in the Gauss-Seidel iterative methods.

2.1 Full-, Half- and Quarter-Sweep Finite Difference Approximation

The five-point Laplacian operators based on the full-, half- and quarter-sweep approaches are shown in Fig. 2 as follows,



(a) (b)



(c)

Fig. 2 The Laplacian operators for (a) full-, (b) half- and (c) quarter-sweep cases.

Therefore, by referring to Fig. 2, the Gauss-Seidel iterative scheme for full-sweep case is defined as [14, 19],

$$q_{i,j}^{(k+1)} \cong \frac{1}{4}\left(q_{i-1,j}^{(k+1)} + q_{i+1,j}^{(k)} + q_{i,j-1}^{(k+1)} + q_{i,j+1}^{(k)} - h^2 s_{i,j}\right) \qquad (3)$$

Half-sweep case [20, 21]:

$$q_{i,j}^{(k+1)} \cong \frac{1}{4}\left(q_{i-1,j-1}^{(k+1)} + q_{i+1,j-1}^{(k+1)} + q_{i-1,j+1}^{(k)} + q_{i+1,j+1}^{(k)} - 2h^2 s_{i,j}\right) \qquad (4)$$

Quarter-sweep case [11-13]:

$$q_{i,j}^{(k+1)} \cong \frac{1}{4}\left(q_{i-2,j}^{(k+1)} + q_{i+2,j}^{(k)} + q_{i,j-2}^{(k+1)} + q_{i,j+2}^{(k)} - 4h^2 s_{i,j}\right) \qquad (5)$$

with $k = 1,2,3,\dots,n$. The solution domain for quarter-sweep approach is shown in Fig. 3, as follows,

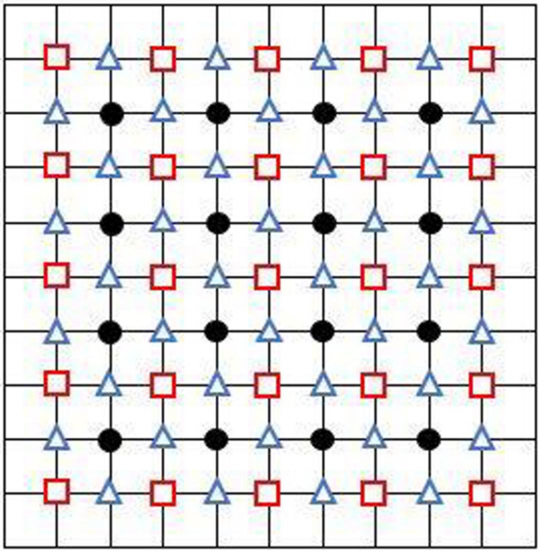


Fig. 3 Solution domain for QSGS iterative method.

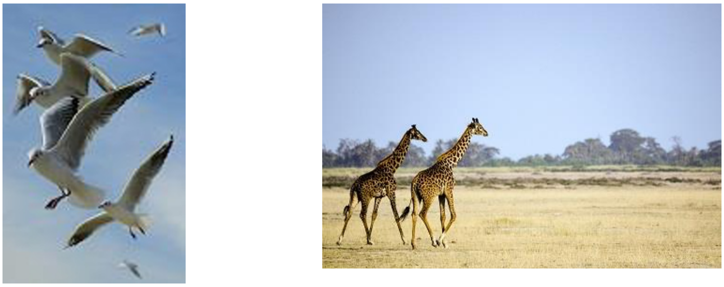
Three types of Laplacian operator are used in the implementation of QSGS method, which are the full-, half- and quarter-sweep operators. The implementation of QSGS method is started by the evaluation of point ●, which is defined in equation (5). Then, the evaluation is followed by the implementation of point □, by using the rotated five-point approximation equation (4) and lastly, the evaluation of point △, by using the standard five-point approximation equation (3).

In this paper, the linear systems formed are solved by using FSGS, HSGS and QSGS iterative methods respectively. Their efficiency based on number of iteration and compositing time are examined and presented in the next section. Basically, the compositing time for half- and quarter-sweep approaches will be faster than the full-sweep approach. This is due to the reason that the computational complexity had reduced about 50% and 75% as compared to full-sweep approach.

3 Numerical Results and Discussion

Three experiment examples were chosen from [22] to carry out the Poisson image blending process. Each examples are in different sizes which comprises of a source and destination images, as shown in Fig. 4.

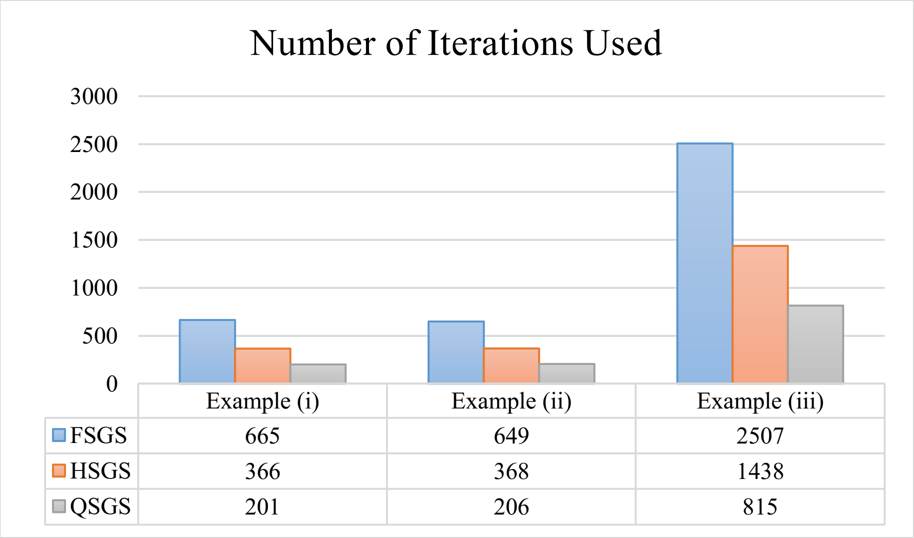
(iii)



(a) (b)

Fig. 4 (a) Source and (b) destination images.

The desired region was selected manually from each source images and then blended into the destination images respectively. The numerical results are shown in Fig. 5 and 6.

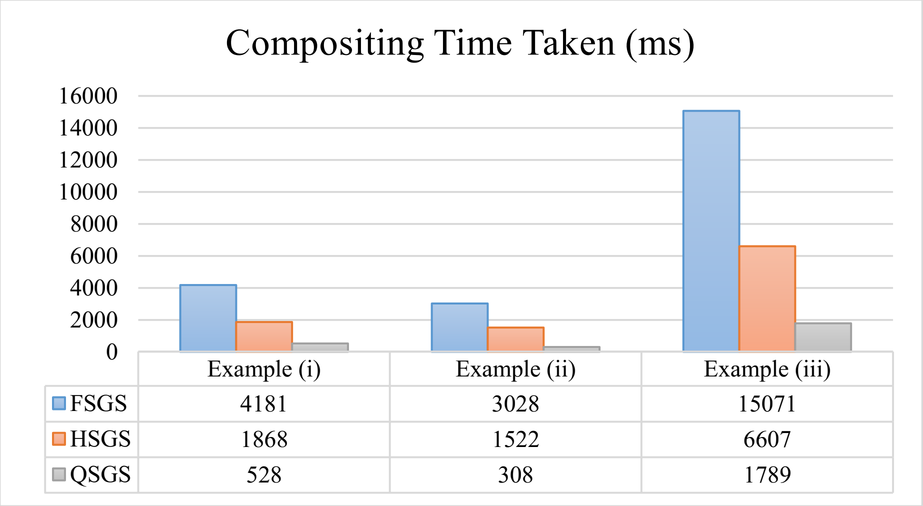


	Example (i)	Example (ii)	Example (iii)
■ FSGS	665	649	2507
■ HSGS	366	368	1438
■ QSGS	201	206	815

Fig. 5 The number of iterations used by the proposed iterative methods.

Fig. 5 above displayed the number of iterations used by each iterative methods to solve the three Poisson image blending problems respectively. The iterative method that used the least number of iterations is the QSGS method, indicated by the grey color bar. Followed by the HSGS method and then the FSGS method. As compared to HSGS and FSGS methods, QSGS method had reduced the number of iterations by approximately 43.43% to 45.08% and 67.49% to 69.77% respectively.

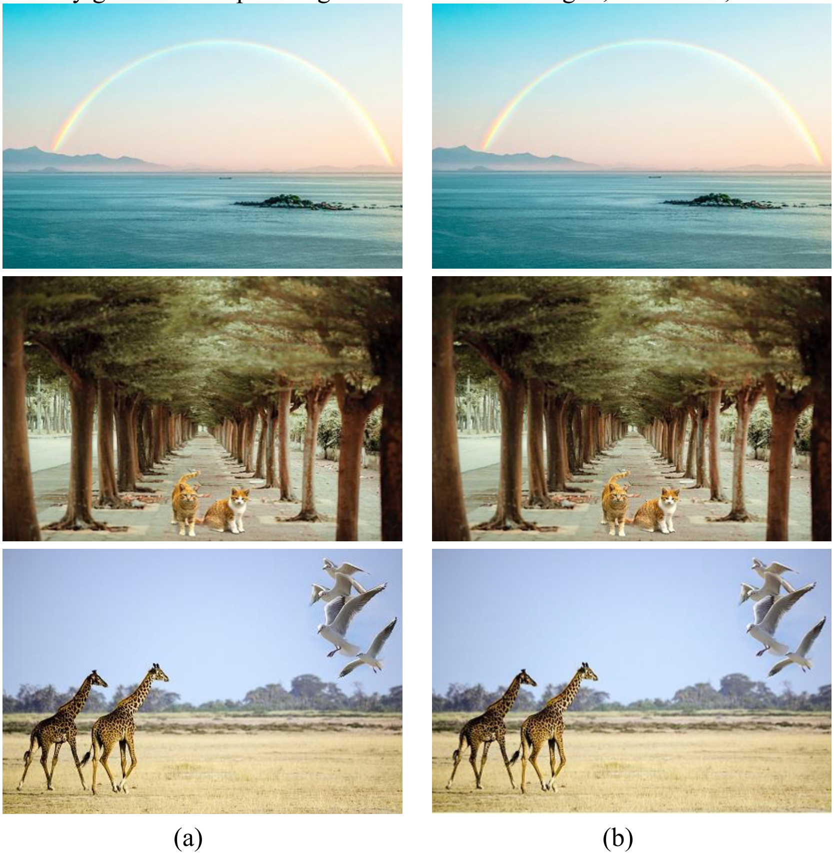
On the other hand, Fig. 6 displayed the compositing time for each examples. It can be seen that the compositing time decreased approximately 87.37% to 89.83% and 49.74% to 56.16% respectively correspond to QSGS and HSGS methods compared to FSGS method.



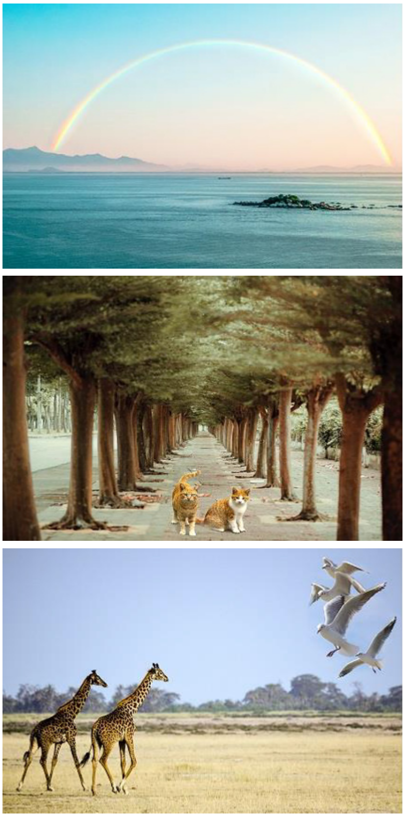
	Example (i)	Example (ii)	Example (iii)
🟦 FSGS	4181	3028	15071
🟧 HSGS	1868	1522	6607
⬜ QSGS	528	308	1789

Fig. 6 The compositing time taken by the proposed iterative methods.

The newly generated output images are illustrated in Fig. 7, as follows,



(a) (b)



(c)

Fig. 7 The new images formed by using (a) FSGS, (b) HSGS and (c) QSGS iterative methods.

By referring to Fig. 7, all the desired region from source images are seamlessly blended into the destination images and formed a natural looking image by using the three proposed iterative methods.

4 Conclusion

From the observation of the numerical results obtained, QSGS method is more superior than the FSGS and HSGS in terms of the number of iterations and compositing time. This is because QSGS method applied the reduction technique which has reduced its computational complexity by approximately 75% as compared to FSGS method. Overall, the newly generated images are with satisfactory visual effect. In future work, we might consider to apply a higher order discretization scheme while obtaining a better result.

References

1. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics 22(3), 313–318 (2003).
2. Qin, C., Wang, S., Zhang, X.: Image editing without color inconsistency using modified poisson equation. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp 397–401. IEEE, Harbin (2008).
3. Morel, J. M., Petro, A. B., Sbert, C.: Fourier implementation of Poisson image editing. Pattern Recognition Letters 33(3), 342–348 (2012).
4. Afifi, M., Hussain, K. F.: MPB: A modified Poisson blending technique. Computational Visual Media 1(4), 331–341 (2015).
5. Hussain, K., Kamel, R. M.: Efficient Poisson Image Editing. ELCVIA Electronic Letters on Computer Vision and Image Analysis 14(2), 45–57 (2015).
6. Wu, H. K., Zheng, S., Zhang, J. G., Huang, K. Q.: GP-GAN: Towards Realistic High-Resolution Image Blending. Computer Vision and Pattern Recognition arXiv:1703.07195 (2017).
7. Srivaramangai, R., Patil, A. S.: Survey of Segmentation Techniques of Cancer Images Emphasizing on MRI Images. International Journal of Computer Science Trends & Technology 3(3), 304–11 (2015).
8. Attia, S. J., Hussein, S. S.: Evaluation of Image Enhancement Techniques of Dental X-Ray Images. Indian Journal of Science and Technology 10(40), (2017).
9. Sabanci, K., Aydin, C.: Using Image Processing and Artificial Neural Networks to Determine Classification Parameters of Olives. Tarım Makinaları Bilimi Dergisi 10(3), 243–246 (2014).
10. Pulido, C., Solaque, L., Velasco, N.: Weed recognition by SVM texture feature classification in outdoor vegetable crop images. Ingeniería e Investigación 37(1), 68–74 (2017).
11. Othman, M., Abdullah, A. R.: An efficient four points modified explicit group Poisson solver. International Journal of Computer Mathematics 76, 203–217 (2000).
12. Sulaiman, J., Othman, M., Hasan, M. K.: MEGSOR iterative scheme for the solution of 2D elliptic PDE's. International Journal of Science, Engineering and Technology 4(2), 264–270 (2010).
13. Ali, N. H. M., Foo, K. P.: Modified Explicit Group AOR methods in the solution of elliptic equations. Applied Mathematical Sciences 6(50), 2465–2480 (2012).
14. Eng, J. H., Saudi, A., Sulaiman, J.: Numerical assessment for Poisson image blending problem using MSOR iteration via five-point Laplacian operator. Journal of Physics: Conference Series 890(1), 012010 (2017).
15. Eng, J. H., Saudi, A., Sulaiman, J.: Numerical Analysis of the Explicit Group Iterative Method for Solving Poisson Image Blending Problem. International Journal of Imaging and Robotics 17(4), 15–24 (2017).
16. Eng, J. H., Saudi, A., Sulaiman, J.: Performance Analysis of the Explicit Decoupled Group Iteration via Five-Point Rotated Laplacian Operator in Solving Poisson Image Blending Problem. Indian Journal of Science and Technology 11(12), (2018).
17. Saudi, A., Sulaiman, J.: Path Planning Simulation using Harmonic Potential Fields through Four Point-EDGSOR Method via 9-Point Laplacian. Jurnal Teknologi 78(8-2), 12–24 (2016).
18. Saudi, A., Sulaiman, J.: Application of Harmonic Functions through Modified SOR (MSOR) Method for Robot Path Planning in Indoor Structured Environment. International Journal of Imaging and Robotics™ 17(3), 77–90 (2017).

19. Eng, J. H., Saudi, A., Sulaiman, J.: Application of SOR Iteration for Poisson Image Blending. In: Proceedings of the International Conference on High Performance Compilation, Computing and Communications, pp. 60–64. ACM, Kuala Lumpur (2017).
20. Abdullah, A. R.: The four point Explicit Decoupled Group (EDG) method: A fast Poisson solver. International Journal of Computer Mathematics 38(1-2), 61–70 (1991).
21. Eng, J. H., Saudi, A., Sulaiman, J.: Implementation of Rotated Five-Point Laplacian Operator for Poisson Image Blending Problem. Advanced Science Letters 24(3), 1727–1731 (2018).
22. Experiments examples are available in https://www.pexels.com

Autonomous Road Potholes Detection on Video

Jia Juang Koh[1], Timothy Tzen Vun Yap[1(✉)], Hu Ng[1], Vik Tor Goh[2], Hau Lee Tong[1], Chiung Ching Ho[1] and Thiam Yong Kuek[3]

[1] Faculty of Computing & Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
[2] Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia
[3]Faculty of Business & Finance, Universiti Tunku Abdul Rahman, 31900 Kampar, Malaysia
jiahuangkoh@gmail.com, {timothy, nghu, vtgoh, hltong, ccho}@mmu.edu.my, kuekty@utar.edu.my

Abstract. This research work explores the possibility of using deep learning to produce an autonomous system for detecting potholes on video to assist in road monitoring and maintenance. Video data of roads was collected using a GoPro camera mounted on a car. Region-based Fully Convolutional Networks (R-FCN) was employed to produce the model to detect potholes from images, and validated on the collected videos. The R-FCN model is able to achieve a Mean Average Precision (MAP) of 89% and a True Positive Rate (TPR) of 89% with no false positive.

Keywords: Road Surface Defects, Object Identification, Video Data, Machine Learning, Deep Learning

1 Introduction

Road defects have been a concern for many drivers as they can cause unnecessary accidents and casualties. The accidents are mainly due to road defects such as potholes, sunken or elevated manholes which are extremely common in many big cities or rural roads. Potholes may cause damage to vehicles such as flat tires, torn-off bumpers, bent wheel rims, and damaged shock absorbers. These defects that could have been the cause of many accidents and they can be avoided if the authorities such as the local councils can be quickly notified for repair.

Nevertheless, the degeneration of road is unavoidable because of constant usage and poor weather conditions. In Malaysia, the government has constantly spent a great deal of money to improve Malaysian roads. For instance, the Selangor state government has spent over half a billion ringgits on improving Selangor roads in 2014 [1]. However, allocation of resources for road maintenance proves to be a challenge. It is costly and time consuming for councils to constantly monitor conditions of roads. Thus, this research seeks to create a model to identify road potholes on video to assist in the maintenance of roads, perhaps aided by autonomous drones that monitor roads in future.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_14

2 Literature Review

In terms of road surface defects detection, primarily potholes, cracks and patches, a few research has been performed for image and video. In 2016, Shen et al., have developed road crack recognition application using MATLAB [2]. The software he developed had successfully extracted a distinct road crack feature from images by employing threshold segmentation and edge detection.

Huidrom et al., proposed a Critical Distress Detection, Measurement and Classification (CDDMC) algorithm for automated detection and measurement of potholes, crack, and patches of a series of road surface condition video frames [3]. The CDDMC algorithm had successfully detected and measured these three specific road surface conditions effectively and precisely in one pass.

Kawai et al., proposed a distinction technique for night-time road surface conditions employing a car-mounted camera [4]. They concentrated on the dissimilarity of features in road surfaces condition. Sun et al., proposed a road image status detection technique using a video camera and developed a Naïve Bayesian classifier to classify the road surface condition image [5].

Zhao-zheng et al., proposed a method to estimate visibility distance based on the contrast of road surface condition with distance information using traffic video-surveillance system [6], while Raj et al., developed an algorithm that detects road surface types such as asphalt, cement, sandy, grassy, and rough based on video data taken from a car-mounted camera [7]. These researches employed image processing techniques with machine learning on video, however, none have applied deep learning approaches, which is investigated in this research work.

3 Methodology

The dataset of images used for the development of the detection model is provided by Nienaber et al., for their research work in South Africa [8, 9]. These images were captured by a GoPro Hero 3+ camera in a vehicle travelling at roughly 40 km/h. Each image has its resolution set to 3680×2760 in JPG format. The dataset contains two different sets, one is considered to be simple (easily recognizable potholes) while another is more complex, their file sizes are 10.8 GB and 16.4 GB respectively. Each of the set consist of folders containing the training images as well as a collection of positive test images. The training images consist of positive data (images that contains potholes) and negative data (images without potholes).

Dataset of on road videos for validation were collected using a GoPro Hero 4 Silver camera mounted on the front of the car. Each video has its resolution set to 1920×1080 and its frame rate set to 30 frames per second. Every video of the dataset is in MP4 format. The entire dataset consists of videos of sudden stops, potholes, smooth roads, speed bumps, uneven roads, corner roads, and rumble strips, with a total file size of 55.4 GB.

The TensorFlow Object Detection API [10] was employed to develop the identification model. TensorFlow [11] is an open source library for deep learning developed by Google. Region-based Fully Convolutional Networks (R-FCN) was chosen as the training model for its precise and highly effective object detection capability. It uses position-sensitive score maps to cope with the dilemma of translation-invariance and translation-variance in image classification and object detection respectively [12].

The performance evaluators employed in this research include the Mean Average Precision (MAP), True Positive Rate (TPR), and False Positive Rate (FPR). The MAP is the percentage form of cases where the potholes were detected over all tested cases, given by

$$MAP = \frac{T_p}{T} \times 100\% \tag{1}$$

where T_p is the number of cases that where the object is detected and T is the total number of tested cases.

The TPR is the percentage form of cases that are detected as positive over all positive cases given by

$$TPR = \frac{T_p}{P} \times 100\% \tag{2}$$

where T_p is the number of cases that the object is detected and P is the number of positive tested cases.

The FPR is the percentage of cases that the object is detected wrongly as positive cases over all negative cases given by

$$FPR = \frac{F_p}{N} \times 100\% \tag{3}$$

where F_p is the number of cases with objects wrongly detected as positive cases and N is the number of negative tested cases.

4 Design of Experiments

A set of images that contained potholes was constructed as the positive data. 1000 positive images were selected from the positive dataset to be labeled. A tool, LabelImg [13], was used to annotate the potholes in the images. A XML file was produced for each labeled image, containing data such as coordinates of the bounding boxes, and their height and width. Correspondingly, a set of 1000 images without any potholes were selected as negative images. Fig. 1 shows samples of the positive (top) and negative images (bottom).

The two sets of positive and negative images were further divided into training and testing sets. The training set contained 90% of all positive and negative data respectively, while 10% of the images were chosen from both positive and negative data respectively to make up the testing set.

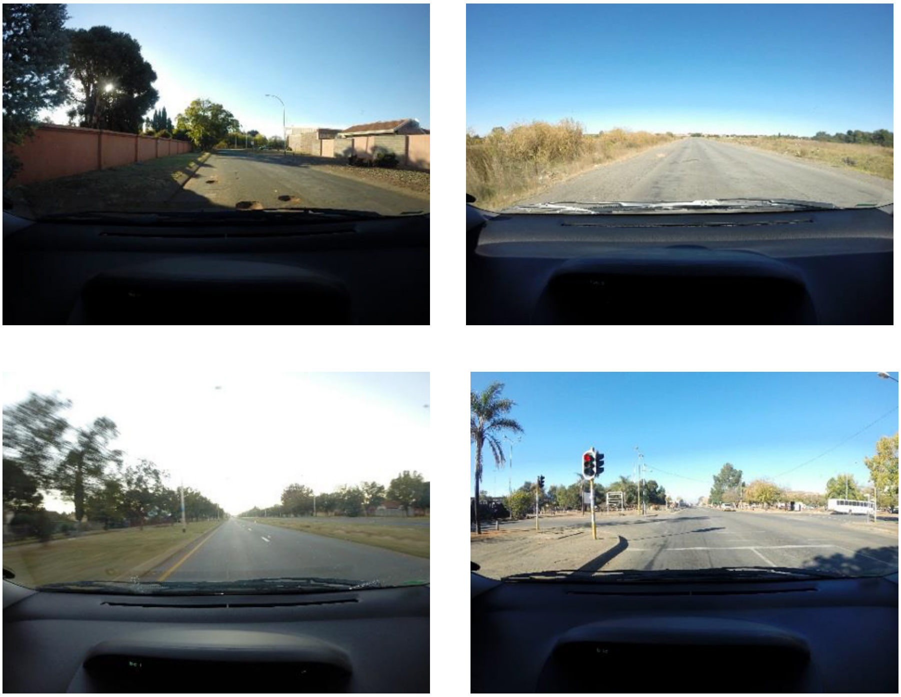


Fig. 1. Sample images of the training and testing set. Top: positive; bottom: negative.

The XML files of both the training data and validation data were converted into a TFRecord, the data object for labeled training data used by the TensorFlow Object Detection API. In addition, a pbtxt file (textual representation of the TensorFlow graph) was also created as the label map.

Training was initiated using the pre-trained models of the R-FCN and its checkpoint alongside the TFRecord of the training data, as well as the label map. Default paramaters (learning rate, number of layers, number of neurons, number of iterations, Lambda L2-regularization parameter, etc) were used, with softmax as the activation function and an epoch of 1. Parameter optimization will be considered in future research work.

The post-trained model of the R-FCN was then used to export the inference graph, from which the final model was generated. The final model was then applied to the test set to obtain the accuracy result. Finally, the model was validated using the videos taken and the detected potholes in the videos were labeled with bounding boxes for visual inspection. A flowchart depicting the process is shown in Fig. 2.

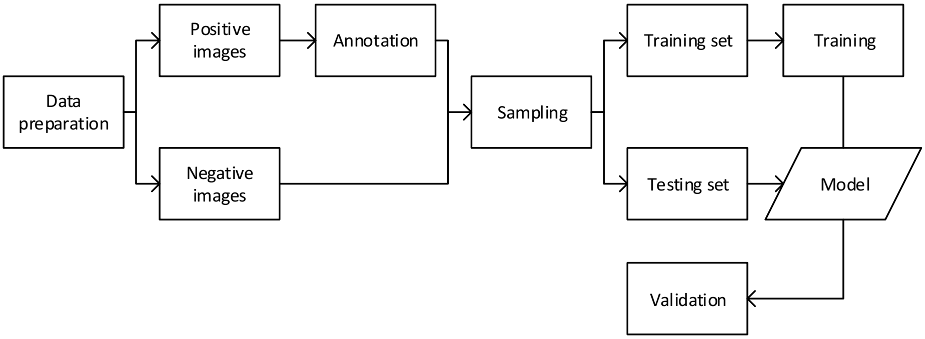


Fig. 2. Flowchart of the experiment process.

5 Results and Discussions

The results of the performance measurement of the model is shown in Table 1. The model is able to achieve a MAP of 89%, with a TPR of 89% and no false positive. In addition, Fig. 3 shows sample results of positive images from the model. The R-FCN can detect most medium sized pothole in bright image successfully, but it has difficulty in detecting potholes in dark or unilluminated areas of the image (Fig.3 – top, right). It also fails to detect small potholes (Fig. 3 – bottom, right).

The R-FCN is able to produce a model with fairly accurate capabilities to detect potholes, but in order to achieve this result, sufficient labeled data has to be available in addition to computing power for continuous training.

Table 1. The MAP and TPR of the model.

MAP	TPR
89%	89%

6 Conclusions

R-FCN was employed using TensorFlow to train a model from images to detect potholes in videos. The training was successful and the R-FCN model is able to achieve 89% MAP and 89% TPR with no false positive. With sufficient labeled data and computational power, R-FCN can achieve high accuracy on pothole detection for use in videos, with limitations in detecting small potholes or potholes in dark or unilluminated areas.

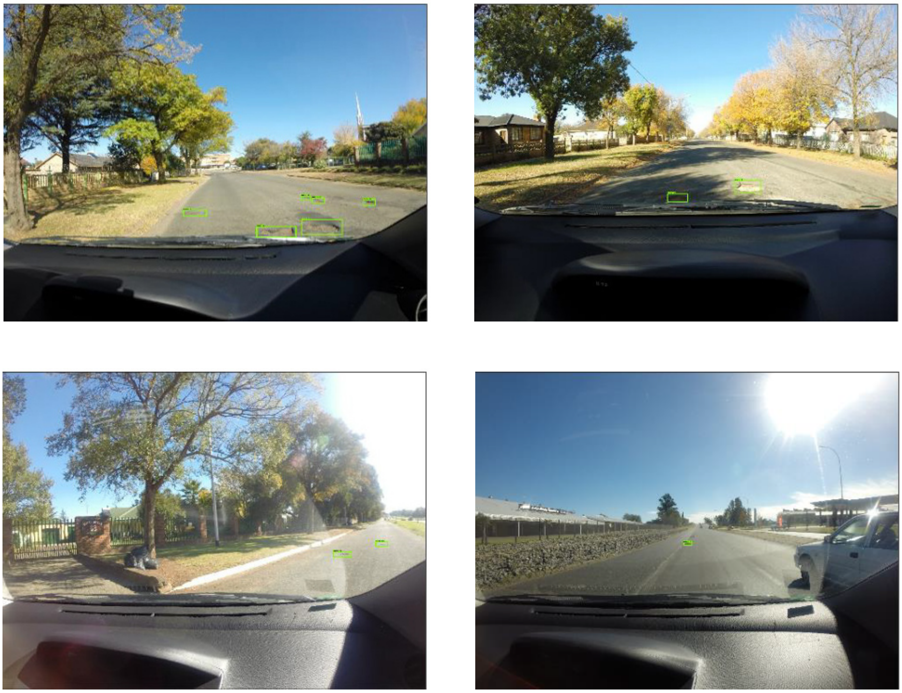


Fig. 3. Positive images with correctly identified potholes.

Acknowledgements. Financial support from the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme with grant number FRGS/1/2015/SG07/MMU/02/1, as well as the Multimedia University Capex Fund with Project ID MMUI/CAPEX170008, are gratefully acknowledged.

References

1. The Star Online, https://www.thestar.com.my/news/community/2014/03/10/crumbling-roads-exhausting-funds-huge-allocation-comes-with-hopes-of-better-maintenance, last accessed 2018/2/10.
2. Shen, G.: Road crack detection based on video image processing. 3RD INTERNATIONAL CONFERENCE ON SYSTEMS AND INFORMATICS 2016, pp. 912–917, IEEE, Shanghai (2016).
3. Huidrom, L., Das, L. K., Sud, S.: Method for automated assessment of potholes, cracks and patches from road surface video clips. Procedia - Social and Behavioral Sciences, 312–321 (2013).
4. Kawai, S., Takeuchi, K., Shibata, K., Horita, Y.: A method to distinguish road surface conditions for car-mounted camera images at night-time. 12TH INTERNATIONAL CONFERENCE ON ITS TELECOMMUNICATIONS 2012, pp. 668–672, IEEE (2012).

5. Sun, Z., Jia, K.: Road surface condition classification based on color and texture information, 9TH INTERNATIONAL CONFERENCE ON INTELLIGENT INFORMATION HIDING AND MULTIMEDIA SIGNAL PROCESSING, pp. 137−140), IEEE (2013).
6. Zhao-Zheng, C., Jia, L., Qi-Mei, C.: Real-time video detection of road visibility conditions. WORLD CONGRESS ON COMPUTER SCIENCE AND INFORMATION ENGINEERING 2009, pp. 472–476. IEEE (2009).
7. Raj, A., Krishna, D., Priya, H., Shantanu, K., Devi, N.: Vision based road surface detection for automotive systems. 2012 INTERNATIONAL CONFERENCE ON APPLIED ELECTRONICS, pp. 223–228, IEEE (2012).
8. Nienaber, S., Booysen, M.J., Kroon, R.S.: Detecting potholes using simple image processing techniques and real-world footage. SATC 2015, Pretoria, South Africa (2015).
9. Nienaber, S., Kroon, R.S., Booysen M.J.: A comparison of low-cost monocular vision techniques for pothole distance estimation. IEEE CIVTS 2015, IEEE, Cape Town, South Africa (2015).
10. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. CVPR 2017 (2017).
11. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC], (2016).
12. LabelImg, https://github.com/tzutalin/labelImg, last accessed 2018/2/10.
13. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. arXiv:1605.06409 [cs.CV], (2016).

Performance Comparison of Sequential and Cooperative Integer Programming Search Methodologies in Solving Curriculum-Based University Course Timetabling Problems (CB-UCT)

Mansour Hassani Abdalla, Joe Henry Obit, Rayner Alfred and Jetol Bolongkikit

Knowledge Technology Research Unit, Universiti Malaysia Sabah,
88400 Kota Kinabalu, Malaysia
mansourabdalla22@gmail.com, joehenry@ums.edu.my,
ralfred@ums.edu.my, jetol@ums.edu.my

Abstract. The current study presents Integer Programming (IP) search methodology approaches for solving Curriculum-Based University Course Timetabling problem (CB-UCT) on real-life problem instances. The problem is applied in University Malaysia Sabah, Labuan International Campus Labuan (UMSLIC). This research involves implementing pure 0-1 IP and further incorporates IP into a distributed Multi-Agent System (MAS) in which a central agent coordinates various cooperative IP agents by sharing the best part of the solutions and direct the IP agents towards more promising search space and hence improve a common global list of the solutions. The objectives are to find applicable solutions and compare the performance of sequential and cooperative IP search methodology implementations for solving real-life CB-UCT in UMSLIC. The results demonstrate both sequential and parallel implementation search methodologies are able to generate and improve the solutions impressively, however, the results clearly show that cooperative search that combines the strength of integer programming outperforms the performance of a standalone counterpart in UMSLIC instances.

Keywords: Timetabling, Integer Programming, Multi-Agent System.

1 Introduction

Timetabling is one of the problems on which so many researches have been done over the years and CB-UCT is an NP-hard and also highly constrained combinatorial problems. This is because specific circumstances give origin to several problems, with an abundance of varying features (constraints) [1]. Timetabling problems are very hard to solve because of these arising problems (varying constraints in every semester). Replicating previous timetable and manually trying to fix the new problem does not solve the problem. In fact, it becomes a burden to academic departments who are involved in timetable generation in every semester.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_15

Timetabling involves two categories of constraints, hard and soft constraints. Hard constraints are mandatory to be satisfied for the timetable to be considered feasible all hard constraints must be fully satisfied, however, soft constraints are not only desirable but the more soft constraints are solved the higher the quality of the solution. In particular soft constraints are used to measure the quality of the timetables. Each institution has their own constraints, some constraints may be considered hard in some institution while some constraints are considered soft in other institution. In addition, the constraints vary from time to time. Additionally, according to [3] modularity is other features which contribute to the hardness of the problems i.e. students are allowed to choose the course from other departments or even from another faculty. Hence to solve all these problems an effective research and search methodology is highly required in this particular domain. In fact, there are so many different techniques proposed in literature such as cooperative search inspired by particle swarm optimisation [2], parallel meta-heuristics [16], parallel local search [2], Parallel Constraint Programming [5] and many more. In recent years scholars acknowledge parallel search as a natural effective solution approach to timetabling problems [8]. However, the questions are what is the best parallel strategy? Can these strategies improve the performance of stand-alone algorithms (i.e. IP, heuristics, and meta-heuristics e.c.t)? Is the proposed parallel IP able to improve the solutions as compared to standalone IP? In order to provide some understanding into these questions, we propose cooperative IP search methodologies to provide some insight into these questions.

In this research, we aim to investigate the performance of sequential and parallel IP for the CB-UCT in UMSLIC. In particular, the standalone sequential IP and parallel IP in which different improving IP agents are running concurrently in different simulated Multi-agent systems are implemented and tested on the real word problem instances. As [2] proposed cooperative search inspired by particle swarm optimisation, this research is inspired by three important objects. Firstly the availability of high performance computer which makes feasible for us to implement IP model which is proven in literature that it requires high performance machine [3], secondly the multi-processor computers and the rise of MAS which motivate parallel processing [4], and finally even though recently there a lot of research devoted to parallel search nonetheless there is little which have taken the advantage of the strength of IP into cooperative search methodology.

Hence the current work contributes to the body of knowledge hereby by proposing both sequential and cooperative integer programming for solving CB-UCT on UMSLIC instances.

2 Related Work

Curriculum-based University Course Timetabling problem (CB_UCT) is very important to research due to their direct importance and relevance in real-life situations [10]. According to [2, 3, 11] requirements differ from one institution to another for any given semester and in fact, according to [11] it is very difficult to produce the general methodology to solve all the problems in every institution. As highlighted by [12] “the problem becomes more complex if the events vary in duration, and each event must occupy only one room for the entirety of this duration”. And so in UMSLIC the problems become complex since the duration of each course are not same as some courses take duration of two hours i.e. main courses, some take three hours i.e. language courses.

Most literature proposes purely heuristic, meta-heuristics [13, 14] and hyper-heuristics solution methods [12]. However, in recent years, integer programming (IP) methods have been the subject of increased attention [12] because of the availability of powerful computers, proven success strength of the IP, and ability to solve large instance in a small amount of time. In addition, cooperative search (Multi-Agent Systems) appears to attract scholars from both artificial intelligence and operational research community [6]. This is because in multi-agent systems-based approaches, intensification and diversification are possible to achieved through agents’ communication and cooperation [2], negotiation of agents to remove the constraints of the event and ability to resources sharing with each other [6], and finally the tendency of guiding algorithms towards more promising search space [3, 4, 8].

However, it is worth noting that, approaches based on operational research do not have good efficiency in solving scheduling problems [6]. Somewhat, they do have easier implementation since they are mostly analyzed by software integrated with efficient and heuristic algorithms [6]. In recent years significant advancement of meta-heuristics in solving university timetabling problems and other complex combinatorial optimization problems have been achieved. These advancements have led to the successful deployment of meta-heuristics in the wide range of combinatorial problems. However, one major drawback of this family of techniques is the lack of robustness on a wide variety of problem instances [15]. Also, the computation times associated with the exploration of the solution space may be very large [8]. Moreover, [16] also emphasized the fact that, the performance of meta-heuristics often depends on the particular problem setting and data.

3 Problem Statement

In every semester, academic institutions are facing difficulties in constructing course timetable. The task is to allocate the set of courses offered by the university to a given set of time periods and available classrooms in such a way no curriculum, lecturer or classroom is used more than once. Essentially the problem in UMSLIC involves assigning a set of 35 timeslots (seven days, with five fixed timeslot per day) according to UMSLIC teaching guidelines.

Each lecturer teaching several courses in each semester and each course has at least one lecture of minimum two hours per week. In addition, UMSLIC’s administration has a guideline for the compulsory, elective, center for promotion of knowledge and language learning (PPIB), and center for co-curriculum and student development (PKPP) courses to be enrolled by the students in each of the semesters throughout the students’ university days Our approach will also fulfill university teaching guideline where there are some general preferences such as some courses particularly program and faculty courses cannot be scheduled on weekends and must be scheduled on the first or third timeslots of the weekdays. In addition, some courses such as PKPP courses cannot take place on weekdays. In addition, some course such as PPIB courses must be scheduled on second, fourth, or fifth timeslot. Hence, this research concentrates on real-life CB-UTT. In fact, in CB-UTT there are five variables identified namely periods, courses, lecturers, rooms, and curricula. The objective is to assign a period and a room to all lectures of each course according to the hard and soft constraints based on UMSLIC teaching guidelines.

4 Sequential IP

The proposed sequential IP is formulated to solve the problem in two stages; in the first stage, the model solves hard constraints. In the second stage, the model tries to deal with soft constraints and maintain the feasibility of the solution. The objective is to generate feasible timetable solution that is able to satisfy all the people affected by the timetable [7].

In particulars, in the first stage, the IP formulation tackle the hard constraints while in the second stage the timetable is being improved by minimizing the soft constraints as much as possible. Firstly the proposed search IP starts with an empty timetable. Since at the start all the problem instances are already pushed to the hash set (list) from the source pre-processed text file and all the information are now in places, the search IP generates room, day, period, and course at random. Then check for feasibility i.e. the room capacity can accommodate the number of students registered in that particular course, the period if is already occupied on that particular day, and if the course is already scheduled. If all the conditions are feasible, then course will be inserted into timetable and the room is registered as already used at that period in a given day. The course is registered as already scheduled, the timeslot is registered as already used and the number of unscheduled course is decremented. The process repeats until all the courses are feasibly scheduled.

First stage algorithm

```
while all courses is not scheduled do
    Select randomly $d \in D$, $r \in R$, $p \in P$, $c \in C$ from the problem instance
if $c \in C$ feasibly can be scheduled i.e. no hard constraint violation then
    Insert into timetable and update the number of course scheduled
    iter++;
else
    Remove $c \in C$ from timetable and update number of course scheduled
end if
end while
```

In the second stage, the simple local search is introduced. The local search gradually tries to improve the quality of the solution generated in the first stage with the knowledge of maintaining the feasibility of the solution. The search is basically based on swapping of events as explained hereby. In this stage, there is two moves, the first move the course is selected randomly and places it into feasible timeslot and room which are also selected at random. In the second move the event is selected at random and insert into empty timeslot. If the course is inserted and the cost factor is improved or even similar to previous cost value the timetable is updated however if the course is inserted but the cost factor is not improved the timetable is not updated. The process keeps repeating until the stopping condition is met which is 300 seconds in this case. The time given is only five minutes; however, the larger the problem instance the higher the time is required [7] and the better the solution.

Second stage algorithm

```
Stop condition: is set 300 seconds
Best solution = initial solution
while stop condition is not met do
    Select two events ($d \in D$, $r \in R$, $p \in P$, $c \in C$) AND ($d' \in D'$, $r' \in R'$, $p' \in P'$, $c' \in C'$) randomly from the feasible solution and swaps them: new solution S*, OR
    Select Event randomly from feasible timetable and insert in to empty slot: new solution S*
    if cost Function(S*) < cost Function (best solution
        Best solution = S*
        iter++;
    else
        Best solution = Best solution
    end if
end while
```

5 Cooperative IP

Figure 1 presents the proposed agent-based IP searches framework. In this research, a decentralized agent-based framework, which consist of given number of agents (*n*) is proposed. Basically this framework is a generic communication protocol for IP search methodology to share solutions among each other. Each IP is an autonomous agent with its own representation of the search environment. All IP agents at the beginning share the same complete feasible solution and then starts with their own search towards more promising search space. Moreover, the communication or ability of the agent to exchange the solutions with one another via the central agents prevent individual agent from trapping on the local optima [8]. Essentially all agents in the distributed environment communicate asynchronously via the central agent. Additionally, it is worth mentioning that, the initial feasible solution is generated by the central agents as well. In clarity this framework will involve asynchronous cooperative communication as follow.

5.1 Central Agent (CA)

The central agent is responsible to generate the initial feasible solution as well as to coordinates the communication process of all other agents involved in the proposed framework. The central agent acts as intermediate agent among the IP agents where it passes the feasible solution and other parameters to the IP agents asynchronously on top of FIPA-ACL communication protocol to improve the solutions. On top of that, the central agent receives the improved solutions from the IP agents and compares the objective function cost value of the received solution with the global solutions on the list, if the improved solution's objective is better than any of the solutions on the global solutions then the worse in the list is replaced. Else the received solution is discarded and the central agent randomly select other solution from the list of the global solutions and send back to that particular agent This procedure continues until the stopping condition is met.

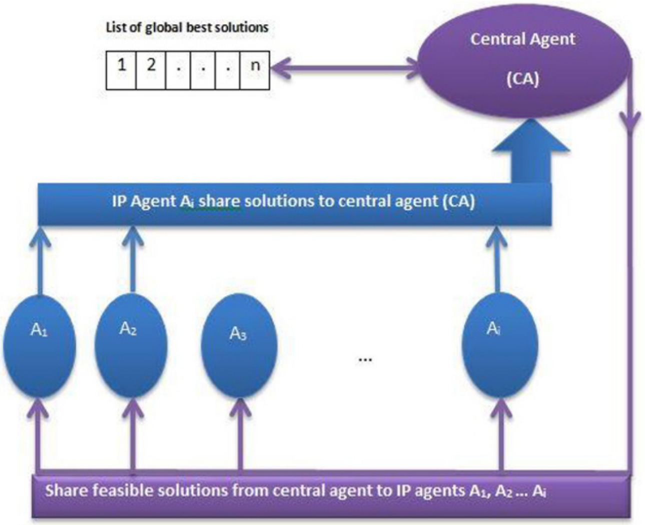

List of global best solutions
1
2
.
.
.
n
Central Agent
(CA)
IP Agent A_i share solutions to central agent (CA)
A1
A2
A3
...
Ai
Share feasible solutions from central agent to IP agents A1, A2 ... Ai

Fig. 1. Proposed Agent-based IP Search methodology Framework

5.2 IP Agents (A_i)

All other agents' start from the complete solution received randomly from central agent and iteratively perform search to improve the solution autonomously. In this case the agents have to maintain the feasibility of the solution i.e. do not violent hard constraint. After certain number of iterations according to the rules stated (after every

30 seconds) the agent passes the solution back to the central agent and request new solution from the central agent. The central agent accepts the solution if only the solution is better to the existing global solutions in the list of the solutions (i.e. the solution objective cost is less than the existing global solutions in the list) else the solution is discarded. If the solution is accepted then the solution with higher objective cost function i.e. worse in the list will be replaced. The reason an improving agent's exchange solution is to make sure the agents are not stuck on local optima, moreover scholars highlighted in the literature that, by exchanging the solution the possibility of the agents (algorithms) changing the position towards more promising space is increased [4, 8, 9].

Best solution Criteria. All of our agents are incorporated with integer programming search methodology. Each agent also is capable to compute the final objective function and return it along with the improved solution. The central agent places all the solutions obtained in a sorted list where the solution on top will be the best solution (the solution with minimum objective function value).

In this framework the value of the objective functions is used to determine the quality of the solution. The lower the cost value the better the solution. Hence for the solution which has improved by the IP agents to be considered better to the global existing solutions, the returned improved solution's objective function should be lower to the one of the available in the global solutions objective functions values. Else the solution is discarded.

6 Experimental Setup and Results

In order to evaluate the performance of the proposed agent-based framework compare to stand alone sequential integer programming we have conducted experiments hereby. The experiments have been carried out using real-life instances on UMSLIC semester one 2016/2017 and semester two 2016/2017. Table 1 gives an overview of the instance characteristics.

Table 1. Summary of the dataset from UMSICL academic division

	Semester1 s2016/2017	Semester2 S2016/2017
Number of student	2263	2224
Number of curriculum	65	49
Number of lectures	108	92
Number of courses	134	117
Cumulative number of constraints	4126	2918

The numbers of unavailability (constraints) in each semester greatly differs between the instances. For example, semester one session 2016/2017 there are 4126 sum of all the hard constraints identified while in second-semester session 2016/2017 there are 2918 sum of constraints identified. It should be noted that the constraints mentioned here refers to the total number of hard constraints for all the courses offered in that particular semester. It can be seen that even it is the dataset from the same university but the number of constraints is not the same for particular two semesters. And this is demonstrated why it is very difficult for the university to duplicate the previous timetables since in every semester the constraints are not the same.

The IP agents implemented in the framework to solve CB-UCT are described in section 4. The central agent reads in the problems and generates initial feasible solutions. The central agent then sends the complete feasible solutions to IP agents to improve the solution. When the search is complete (complete n search) the central agent receives the results from the improving agents and insert into the sorted list of size *n*. However, after the list is already with n different solutions, whenever the central agents receive the new solution from the IP agents it will compare its objectives with the existing solutions in the list as explained in section 5 of this paper.

As described in research objectives the tests are designed to compare different groups of IP agents with their Stand-Alone (SA-IP) counterparts. For each scenario, the experiments were conducted over 50 runs for each problem instance and the average objective values were computed. Basically, we conducted 50 runs for each problem instances because we wanted to find the upper and lower bound and hence the overall consistence of the algorithms. The agents conducted only 30 messages to complete each search taking no longer than five minutes to complete whole experiments. The number of conversations 30 is chosen because experimentation shows that the rate of solution improvement is reduced after that number. The 30 conversations last no longer than about five minutes and this is deemed to be a good stopping condition. The results are shown in table 2.

Table 2. Experimental results for the proposed Standalone IP (SA-IP) and Cooperative IP search.

	Number of agents	Semester1 s2016/2017	Semester2 s2016/2017
Initial cost	0-1 IP	368.04	377.29
Final Average cost	SA-IP	326.94	343.69
Final average cost	3	321.20	338.80
Final average cost	6	302.20	318.50
Average improvements (%)	SA-IP	10.99	8.91
Average improvements (%)	3	12.73	10.20
Average improvements (%)	6	17.89	15.58

The improvement from the initial to final cost value for the Standalone IP (SA-IP) is 10.99 and 8.91 for s1 2016/2017 and s2 2016/2017 respectively. Also when three IP agents (Ai) are used is 12.73% and 10.20% for s1 2016/2017 and s2 2016/2017 respectively. On the other hand, the improvement of the solution's cost value when six IP agents (Ai) is used is 17.89% and 15.58% % for s1 2016/2017 and s2 2016/2017 respectively.

The results presented clearly demonstrate that cooperative search outperforms standalone IP in this context. The IP agents improve solution as compared to standalone IP. In addition it worth to note that there is huge possibility that, the solution can be improved even further. Because in fact in this experiments it is not purely parallel as this is just simulation of parallel computing hence the performance might increase more as if each agent run on the self-machines.

The main benefits of the agent-based approach adopted for CB-UCT in UMSLIC are the possibilities of intensifying and diversifying the search space, where IP agents (*Ai)* are able to changes solutions with each other in the distributed MAS [2]. This leads the improving agents to easily move towards the most promising search areas of the search space [6]. Basically, by the analysis, the results, the numbers of IP agents used in the framework determine the quality of the solution generated. In this regard, we find that the quality of the solution in this framework proves to increase slightly as the number of IP agents (Ai) is increased.

7 Conclusion

In this research, we have conducted a comprehensive study of sequential IP search methodology approach in solving CB-UCT. In addition, we have focused on methods based on a cooperative search by incorporating sequential IP into agent-based multi-agent systems. In the current study, we have demonstrated on how IP can be integrated into MAS in order to conduct the cooperative search in solving the CB-UCT. To prove this hypothesis, we have justified the capabilities of MAS and how cooperative search can be natural approaches to solve the problem and find higher quality solutions as compared to the standalone sequential counterpart.

The advantages of using MAS in CB-UCT as compared to standalone IP in this context is the ability for the IP agents to share the best part of the solutions and the possibility of the agent moving towards more promising search space.

In general, cooperative outperform standalone IP can be attributed to the fact in standalone IP once the algorithms stuck on local optima it cannot improve solution anymore however in cooperative search once agent stack on local optima agent can changes solutions and through that, the agent is able to escape from local optima.

References

1. Landir S, Maristela O.S., Alysson M.C.: Parallel local search algorithms for high school timetabling problems, European Journal of Operational Research,Volume 265, Issue 1, 2018, Pages 81-98, ISSN 0377-2217,
2. Obit, J. H., Alfred. R., Abdalla, M.H.: A PSO Inspired Asynchronous Cooperative Distributed Hyper-Heuristic for Course Timetabling Problems. Advanced Science Letters, (2017)11016-11022(7)
3. Obit. J. H., Ouelhadj, D., Landa-Silva, D., Vun, T. K.., Alfred, R.: Designing a multi-agent approach system for distributed course timetabling. IEEE Hybrid Intelligent Systems (HIS), 10.1109/HIS(2011)-6122088.

4. Lach, G., & Lübbecke, M. E. (2012).: Curriculum based course timetabling: new solutions to Udine benchmark instances. Annals of Operations Research, 194(1), 255-272
5. Regin JC, Malapert A.: Parallel Constraint Programming. (2018). springerprofessional.de. Retrieved 17 April 2018.
6. Babaei, H., Hadidi, A A.: Review of Distributed Multi-Agent Systems Approach to Solve University Course Timetabling Problem. Advances In Computer Science : An International Journal, 3(5), 19-28. (2014).
7. Lach, G., & Lübbecke, M. E.: Curriculum based course timetabling: new solutions to Udine benchmark instances. Annals of Operations Research, 194(1), 255-272 (2012).
8. Cung, V.-D., Martins, S. L., Ribeiro, C. C., Roucairol, C.: Strategies for the parallel implementation of metaheuristics. In Essays and surveys in metaheuristics (pp. 263-308): Springer (2002).
9. Obit, J. H.: Developing novel meta-heuristic, hyper-heuristic and cooperative search for course timetabling problems. Ph.D. Thesis, School of Computer Science University of Nottingham (2010)
10. Babaei, H., Karimpour, J., & Hadidi, A. (2015).: A survey of approaches for university course timetabling problem. Computers & Industrial Engineering, 86, 43-59. doi:10.1016/j.cie.2014.11.010
11. Obit. J.H., Yik. J. K., Alfred. R.: Performance Comparison of Linear and Non-Linear Great Deluge Algo...: Ingenta Connect. (2018). Ingentaconnect.com. Retrieved 17 April 2018, from http://www.ingentaconnect.com/content/asp/asl/2017/00000023/00000011/art00129
12. Antony E.P, Hamish W., Matthias E., David M.R, Integer programming methods for large-scale practical classroom assignment problems. (2015). Computers & Operations
13. Yik , J. K., Obit, J. H., Alfred, R.: Comparison of Simulated Annealing and Great Deluge Algorithms for...: Ingenta Connect. (2018). Ingentaconnect.com. Retrieved 17 April 2018,
14. Norgren, E., Jonasson, J.: Investigating a Genetic Algorithm-Simulated Annealing Hybrid Applied to University Course Timetabling Problem: A Comparative Study Between Simulated Annealing Initialized with Genetic Algorithm, Genetic Algorithm and Simulated Annealing. DIVA. Retrieved 17 April 2018,
15. Di Gaspero, L., McCollum, B., Schaerf, A.: The second international timetabling competition (ITC-2007): Curriculum-based course timetabling (track 3).
16. Crainic, T. G., Toulouse, M.: Parallel strategies for meta-heuristics. In Handbook of metaheuristics (pp. 475-513(2003)): Springer.

A Framework for Linear TV Recommendation by Leveraging Implicit Feedback

Abhishek Agarwal[1], Soumita Das[1], Joydeep Das[2] and Subhashis Majumder[1]

[1] Dept. of Computer Sc. & Engg., Heritage Institute of Technology,
Kolkata, WB, India
agarwlabhishek@gmail.com
soumitamum210@gmail.com
subhashis.majumder@heritageit.edu
[2] The Heritage Academy, Kolkata, WB, India
joydeep.das@heritageit.edu

Abstract. The problem with recommending shows/programs on linear TV is the absence of explicit ratings from the user. Unlike video-on-demand and other online media streaming services where explicit ratings can be asked from the user, the linear TV does not support any such option. We have to rely only on the data available from the set top box to generate suitable recommendations for the linear TV viewers. The set top box data typically contains the number of views (frequency) of a particular show by a user as well as the duration of that view. In this paper, we try to leverage the feedback implicitly available from linear TV viewership details to generate explicit ratings, which then can be fed to the existing state-of-the-art recommendation algorithms, in order to provide suitable recommendations to the users. In this work, we assign different weightage to both frequency and duration of each user-show interaction pair, unlike the traditional approach in which either the frequency or the duration is considered individually. Finally, we compare the results of the different recommendation algorithms in order to justify the effectiveness of our proposed approach.

Keywords: Recommender Systems, Linear TV, Collaborative Filtering, Implicit Feedback

1 Introduction

Recommender systems (RS) [1] produce recommendations through algorithms like collaborative filtering (CF) or content-based filtering. Content-based filtering [3] predict preferences based on the content of the items and the interests of the users, while CF [13] builds a model from a user's past behavior (items previously consumed or ratings given to those items) as well as decisions made by other similar users. This model is then used to predict items that the user may have an interest in. CF algorithms are of two types: memory based and model based algorithms. Memory based algorithms identify the top-K most similar users (neighbors) to the active user, and then use a weighted sum of the ratings of the neighbors to predict missing ratings for the active user [3]. Model based algorithms [11], in contrast, implement data mining or machine learning algorithms on the training data to estimate or learn a model to make predictions for an active user. Model based algorithms handle the sparsity and scalability

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_16

problems better than the memory based algorithms. The disadvantages of this technique are in model building and updating that often turn out to be costly.

The large number of TV programs give users many choices, however, it may confuse the users as it is not easy to find a TV program that is interesting, because of the tremendous number of choices available. Thus, TV program recommender systems have become really important. Traditional RS typically use a user-item rating matrix which records the ratings given by different users to different items. However, in case of recommending TV programs we do not get explicit ratings from the users since there is no option for them to rate a particular show on the TV. Therefore we need to leverage the implicit data available (provided by set top boxes) in the system to provide suggestions for TV shows to the existing as well as new users. This implicit feedback focuses mainly on the number of times a particular user has watched a particular show (frequency of user-show interaction) and the corresponding time duration for which the user has watched the show (duration of user-show interaction). Majority of the existing TV recommendation systems [2,7,14] rely only on implicit feedback. However, in order to apply standard CF algorithms, we need to convert the feedback into numeric ratings. In this work, our primary objective is to map the implicit feedback into explicit ratings and then use any of the existing CF based algorithms to generate recommendations.

The previous researches done in the domain of TV recommendation either consider the frequency of user-show interaction [10] or the duration of user-show interaction [14] along with demographic information of the users. Only considering the frequency of views while recommending shows has drawbacks, because it alone cannot indicate whether a user liked or disliked a particular show. For example, if a user switches to a particular show often, it will lead to a higher frequency of views and indicate that the user really likes the show. However, the duration of these views can be very short which might rather indicate that the user does not like the show that much and also does not like to watch it for a long period of time. This kind of observation is very common when a user is surfing through the channels searching for something interesting or during the commercial breaks. Similarly, duration of a view cannot alone indicate whether a user likes or dislikes the show. A higher duration of view may indicate appreciation of one particular episode but a corresponding lower frequency of view may indicate otherwise. Thus, we need to infer the implicit feedback properly by giving appropriate weightage to both the frequency and duration of all such views. In this paper, we propose a recommendation framework where we consider both the frequency and duration of user-item interaction and also assign different weightage to both of them in order to achieve best possible recommendations. The results of the experiments conducted have shown that when both these factors are considered together, the recommendations are more effective than the case when either one of them were considered separately.

The rest of the paper is organized as follows: In section 2, we review some of the past works related to TV recommendation. In section 3, we present the solution framework and our proposed approach. In section 4, we describe our experimental settings and in section 5, we report and interpret our results. We conclude in section 6 discussing our future research directions.

2 Related Work

Recommender systems play a very important role in increasing the popularity of linear TV. Personalized suggestions for TV programs help linear TV compete with the modern video-on-demand services. In one of the early papers on TV recommendation, the authors present the idea of a personalized Electronic Program Guide (EPG) [8]. They identified some important research questions on TV program recommendation including user profiling methods, use of recommendation algorithms and how to use group recommendation to TV users. Another personalized EPG based TV recommendation was proposed [6] where the authors use one hybrid recommendation algorithm to learn users' preferences in terms of different TV channels, genres, etc. to generate new recommendations. Ardissono et al. [2] also proposed a hybrid recommendation approach on the basis of implicit preferences of the users captured in terms of program genres and channels, user classes and viewing history. All these information were gathered from users set top box or were downloaded from satellite stream. The importance of social media in TV shows recommendation is exploited by Chang et al. [5], where the authors propose a user preference learning module that includes user's past viewing experience as well as friendship relations in social networks. Cremonesi et al. [7] proposed a context based TV program recommendation system where the current context of the user along with implicit feedback is explored while making suggestions. There have been some work that intended to compute the top channel, top channel per user and also top channel per user per slot [14]. This is computed based on popularity, which is calculated in terms of total watching minutes accumulated by the channel. They used two important functions namely score aggregation and rank aggregation in order to provide effective recommendations. In this work, we map the implicit preferences of users into explicit ratings and then use standard recommendation algorithms to generate recommendations.

3 Solution Framework

Unlike conventional RS, recommending TV shows is more challenging due to several reasons. Firstly, content of TV programs change over time. Some TV programs are broadcast only once (e.g. movies) and do not repeat over a specific period of time, while some other shows repeat on the same day (e.g. episode of a show). The dynamic content of TV programs become a constraint in providing effective recommendations. Secondly, linear TV programs have a predefined schedule and therefore the set of recommended items is confined to the programs getting broadcasted at the moment when the recommendation is sought. Thirdly, feedback of the users about different TV shows are usually implicit (viewed/not viewed). Therefore it becomes difficult to implement pure CF algorithms to generate recommendations for linear TV since we do not get explicit ratings of the TV shows from the users. Fourthly, it is not possible for a user to watch multiple shows at the same time. Thus any recommendation algorithm must consider the programs which are scheduled simultaneously so that the most interesting shows can be recommended to the users. In this paper, we address the TV shows recommendation problem by analyzing the implicit feedback of the users in order to find the most important features and then provide appropriate weightage to those features. In other words, we try to convert the implicit feedback of a user-program interaction pair into an explicit rating by assigning proper weightage to the different features of the said user-program interaction pair.

3.1 Proposed Approach

Note that the two most important features of any user-show interaction are (a) frequency- the number times a user U has watched a show P over a given period of time, and (b) duration- it is the amount of time a user U spends watching a unique instance of a show P. First, we calculate the frequency of each unique user-show interaction pair. The average duration of view for an unique user-show interaction pair is calculated by summing the durations of each view of the pair and then dividing the sum by the frequency of that pair. Further, let us state two important points regarding the behavior of TV users.
(1) Most users tend to skip advertisements during a break which reduces the actual running time of the show. For example, a show that has been scheduled to run for 30 minutes actually has only 22 minutes of content. The rest 8 minutes might be spent on commercials and other promotional activities during which the users tend to switch channels.
(2) A lot of shows are broadcasted multiple times a day and most of the users have a tendency to watch it only once. Hence, the total number of unique instances of a particular show is crucial.

In order to get better and accurate results the above two points need to be considered before proceeding with any further computations. Since, the information related to the above two points cannot be inferred directly from the data set we are using, we need to make the following assumptions:

- Since the actual running time of a show excluding the commercials is not available to us, we consider the maximum average duration of that show from all the users, and use it as the actual running time (ART) of that show.
- Since an instance (episode) of a show may be broadcasted multiple times, we need to find the actual frequency of that show. For this, we count the number of unique instances of that show and use it as a measure for the total frequency (TF) of that show. An instance of the above process is shown in Table 1 and Table 2.

In Table 1, we can observe that user $U1$ watched the show $P5$ four times having unique event ids $E1, E2, E3$, and $E4$. Accordingly its average duration is 21.5 and frequency is 4 (see Table 2). Similarly we calculate the average duration and frequency for all other users who watched the show $P5$. In Table 2, we compute ART (22) and TF (4) for the show $P5$ by taking the maximum of average duration and frequency respectively.

Table 1: Sample User-Show Interaction

User Id	Program Id	Event Id	Duration of view
U1	P5	E1	22
U2	P5	E1	15
U3	P5	E1	23
U1	P5	E2	23
U2	P5	E2	5
U3	P5	E2	21
U1	P5	E3	22
U2	P5	E4	20
U1	P5	E4	19
U4	P5	E4	5

Table 2: Calculation of *ART* and *TF*

User Id	Program Id	Average Duration	Frequency
U1	P5	21.5	**4**
U2	P5	13.33	3
U3	P5	**22**	2
U4	P5	5	1

3.2 Mapping Implicit Feedback into Explicit Rating

To convert the available implicit feedback into explicit ratings we need to scale the feedback obtained in terms of duration of view and frequency of view. In this work, we define two ratios namely, Duration Ratio (DR) and Frequency Ratio (FR), which will help us to compare the individual implicit feedback with the overall available feedback.

Duration Ratio (DR): It is the ratio of the average duration of view of each unique user-show interaction to the ART of that show. The values will range from 0 to 1.

Frequency Ratio (FR): It is the ratio of the frequency of each unique user-show interaction to the TF of that show. The values will range from 0 to 1.

We calculate the frequency ratio and duration ratio corresponding to each unique user-show interaction. The above ratios will help us to estimate the explicit ratings from the available implicit feedback as follows. The first step is to 'bin' the range of values obtained as DR and FR above. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent but need not be of equal width. In our case, the number of bins will depend on the rating scale (1 - 5). We divide the entire range of values (DR and FR) into a series of intervals and then assign a bin to each of the intervals. We consider the bins for the FR to be of equal width while the bins for the DR to be of unequal width. The width of the bins are decided based on the findings derived from the available dataset. In this work, we limit the number of bins to 5 since we aim to derive ratings on a scale of 1 to 5 from the available implicit feedback. An example of this process is shown in Table 3 and Table 4.

Table 3: Duration Rating

Bin no.	Duration Ratio (DR)	Rating
1	> 0.0 & ≤ 0.05	1
2	> 0.05 & ≤ 0.25	2
3	> 0.25 & ≤ 0.50	3
4	> 0.5 & ≤ 0.75	4
5	> 0.75 & ≤ 1.0	5

Table 4: Frequency Rating

Bin no.	Frequency Ratio (FR)	Rating
1	> 0.0 & ≤ 0.2	1
2	> 0.2 & ≤ 0.4	2
3	> 0.4 & ≤ 0.6	3
4	> 0.6 & ≤ 0.8	4
5	> 0.8 & ≤ 1.0	5

In order to assign ratings to each of the unique user-show interaction, we take help of the binning process discussed above. The corresponding bin number will tell us the derived rating. As for example, in Table 3, if the DR of a user-show interaction falls in the range of $0.01 - 0.05$, then its bin no. is 1 and accordingly assigned a rating of 1. Similarly, a DR in the range of $0.76 - 1.0$ corresponds

Table 5: Final Rating Calculation

User Id	Program Id	DR	FR	$R_{Duration}$	$R_{Frequency}$	R_{Final}
U1	P5	0.97	1	5	5	5
U2	P5	0.6	0.75	4	4	4
U3	P5	1.0	0.5	5	3	3.87
U4	P5	0.22	0.25	2	2	2

to bin no. 5 and as a result a rating of 5. We derive similar ratings using FR as shown in Table 4. Thus, for each unique user-show pair we actually get two kinds of ratings, one rating corresponding to the frequency of the view (obtained from the bins for FR) and another one corresponding to the duration of the view (obtained from the bins for DR). Henceforth, we will refer to these ratings as frequency rating ($R_{frequency}$) and duration rating ($R_{duration}$) respectively. Although we have derived two ratings $R_{frequency}$ and $R_{duration}$ for each user-show pair, our aim is to find a single rating combining the two ratings, since none of the state-of-the-art recommendation algorithms allow multiple ratings for a unique user-item pair. We term this rating as R_{final} and is calculated using the following equation.

$$R_{final} = (R_{frequency})^n \times (R_{duration})^{(1-n)}, \; where \; 0 \leq n \leq 1 \qquad (1)$$

Here n is the weightage of $R_{frequency}$ and $(1-n)$ is the weightage of $R_{duration}$. We determine the value of n experimentally to maximize the accuracy of the final ratings. We will discuss more about it later on in the Experimental section. An example of this rating calculation using $n = 0.5$ is shown in Table 5. Once we obtain the ratings for the different user-show interactions we can use any standard CF based algorithm to produce recommendations. In this work, we have tested our scheme using User-based CF, Item-based CF, SVD, NMF, and PMF methods of recommendation. We have depicted our framework pictorially in Figure 1.

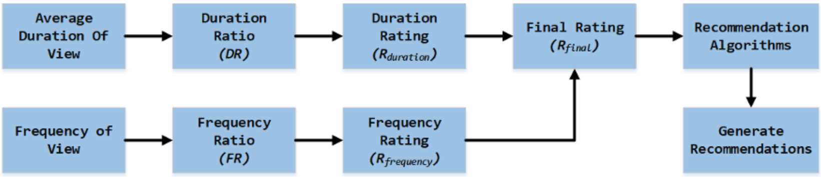

Average Duration Of View
Duration Ratio (DR)
Duration Rating ($R_{duration}$)
Final Rating (R_{final})
Recommendation Algorithms
Frequency of View
Frequency Ratio (FR)
Frequency Rating ($R_{frequency}$)
Generate Recommendations

Fig. 1: Flowchart of our Framework

4 Experimental Settings

4.1 Data Description

We use a dataset containing viewing history of around 13,000 users over 217 channels[3]. The data has been recorded over a period of 12 weeks. There are

[3] http://recsys.deib.polimi.it/?page_id=76

14,000 programs/shows available to the users. The data set consists of information such as user id, program id, channel id, slot and duration of each view. We have considered only those user-program interactions, where the duration of view was greater than one minute. There are about 12.3 million such interactions in our dataset. We have divided the dataset into five disjoint sets. Each set has been used separately for testing and then the rest four for training, so that there were five different training/testing sets. We have repeated our experiment with each set and then considered the average of the results.

4.2 Evaluation Metric Discussion

The prediction accuracy of our algorithm is measured in terms of Root Mean Square Error (RMSE) [13]. The objective of any recommendation algorithm is to minimize the RMSE value. However, only RMSE cannot correctly evaluate a Top-k recommendation list. Therefore in this work, we use Precision, Recall and F1 measure metric [9] to evaluate the quality of the recommended list.

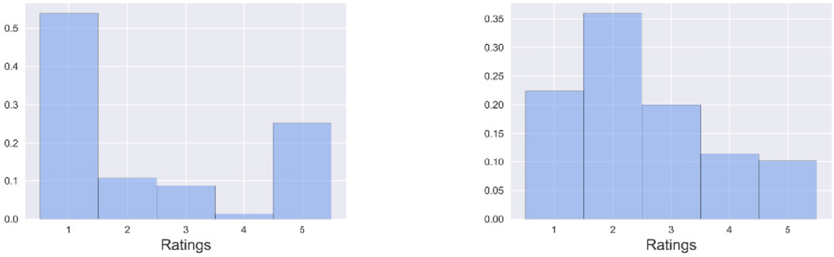
$$Precision = \frac{t_p}{t_p + f_p} \quad Recall = \frac{t_p}{t_p + f_n} \quad F1 = \frac{2 * Precision \ * Recall}{Precision + Recall}$$

True Positive (t_p) or a hit means a relevant product is recommended to a customer by the recommender system. On the contrary, False Positive (f_p) denotes the case when an irrelevant item is recommended, and when an item of customer's liking has not been recommended then we term the case as False Negative (f_n). F1 measure combines Precision and Recall with equal weightage making the comparison of algorithms across datasets easy.

5 Results and Discussion

We report the frequency rating and duration rating distributions in Figure 2(a) and 2(b) respectively. From the distribution reported in Figure 2(a), we can infer that about 54% of the time the frequency rating $R_{frequency}$ is 1. This is due to the fact that a lot of users tend to surf through different channels looking to explore new content but they watch only a handful of the TV shows on a regular basis. On the other hand only around 25% of the time users watched the show on a regular basis, which is indicated by a $R_{frequency}$ value of 5. Similarly from Figure 2(b), we can observe that about 22.5% of the time the users watched the show for less than 5% of the actual running time of the show as indicated by a duration rating $R_{duration}$ of 1. This happens mostly when a user is surfing through the channels looking for new shows. We can further notice that around 10% of the time users watched the show for more than 75% of the actual running time of the show indicated by an $R_{duration}$ value of 5.

From the above observations, we can conclude that the situation for linear TV is quite different from the online streaming services and VOD services like Netflix, Amazon Prime, etc. Users do not have the freedom to watch any show at any given time. Therefore, a lot of users are not able to watch TV shows that frequently (they might not always be available when a particular show gets broadcasted). Moreover, users dont have the freedom to rewind or pause a show



(a) Frequency Rating

(b) Duration Rating

Fig. 2: Normalized Distribution of Ratings

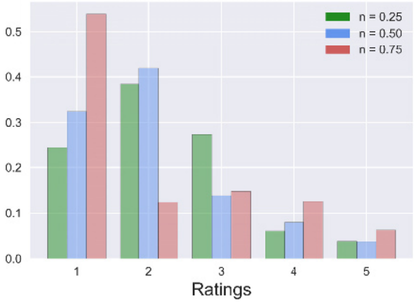


Fig. 3: Normalized Distribution of Final Ratings

whenever they want to. Thus, users may not be able to completely watch a show.

5.1 Finding Right Weightage

A common feature of linear TV viewership is that a lot of users do not watch TV frequently and also tend to watch it for shorter duration. Therefore, depending solely on frequency or duration to make recommendations will be unwise. A balanced approach needs to be taken where the right weightage is assigned to both frequency and duration to make the suitable recommendations. Therefore, in the calculation of R_{final}, we give weightage to both frequency and duration (see equation 1). We vary the value of n from 0 to 1 in order to find the optimum weightage that will maximize the accuracy of the final rating. An example of derived final rating using $n = 0.25, 0.5$, and 0.75 is shown in Figure 3.

5.2 Comparisons

In this work, we used two popular memory based CF algorithms - User-based and Item-based [13], and three matrix factorization techniques namely Singular Value Decomposition (SVD) [11], Non-negative Matrix Factorization (NMF) [4] and Probabilistic Matrix Factorization (PMF) [12]. These recommendation methods are combined with our framework to verify whether their performance has improved or not. We compute user-user and item-item similarities using cosine-based similarity measure. In SVD, the user-item matrix is decomposed into three matrices with n features: $R = U_n S_n V_n^T$. The prediction score for the i-th customer on the j-th product is given by $P_{i,j} = \bar{r_i} + U_n \sqrt{S_n}^T(i).\sqrt{S_n} V_n^T(j)$, where

Table 6: Recommendation Performance Comparisons in terms of RMSE, Precision, Recall and F1. The bold numbers indicate best results

Recommendation Method	n	$(1-n)$	$RMSE$	$Precision$@10	$Recall$@10	$F1$@10
User-Based	0	1	1.093	0.703	0.313	0.433
	0.25	0.75	0.836	0.603	0.324	0.421
	0.5	0.5	0.664	0.768	0.626	0.689
	0.75	**0.25**	**0.51**	**0.985**	**0.692**	**0.812**
	1	0	0.544	0.978	0.645	0.777
Item-Based	0	1	1.065	0.886	0.191	0.314
	0.25	0.75	0.866	0.952	0.087	0.16
	0.5	0.5	0.7	0.879	0.368	0.518
	0.75	**0.25**	**0.568**	**0.992**	**0.587**	**0.737**
	1	0	0.57	0.987	0.569	0.721
SVD	0	1	1.051	0.73	0.322	0.462
	0.25	0.75	0.794	0.65	0.327	0.449
	0.5	0.5	0.613	0.79	0.569	0.68
	0.75	**0.25**	**0.434**	**0.986**	**0.661**	**0.815**
	1	0	0.474	0.979	0.588	0.734
PMF	0	1	1.051	0.739	0.316	0.458
	0.25	0.75	0.797	0.66	0.316	0.442
	0.5	0.5	0.617	0.797	0.554	0.673
	0.75	**0.25**	**0.438**	**0.987**	**0.658**	**0.813**
	1	0	0.478	0.98	0.585	0.733
NMF	0	1	1.071	0.744	0.272	0.413
	0.25	0.75	0.825	0.656	0.258	0.384
	0.5	0.5	0.651	0.773	0.558	0.666
	0.75	**0.25**	**0.485**	**0.981**	**0.662**	**0.814**
	1	0	0.52	0.973	0.593	0.736

$\bar{r_i}$ is the i-th row average. For both SVD and PMF methods, we consider 40 features while NMF is implemented using 15 features.

We report and compare the recommendation performance using different recommendation methods in Table 6. Note that, we present Precision (P@10), Recall (R@10) and F1 ($F1$@10) score on position 10. The bold numbers indicate the best results for that particular recommendation method. In Table 6, n is the weightage of $R_{frequency}$ and $(1-n)$ is the weightage of $R_{duration}$. A study of Table 6 clearly reveals that when $n = 0.75$, we get the best results for all the recommendation methods irrespective of the different evaluation metrics. This indicates that frequency of view is a more significant factor than duration of view in order to generate effective recommendations. We can further notice that when only frequency is considered ($n = 1$) or when only duration is considered ($n = 0$), then the recommendation results are worse than the case when $R_{frequency}$ is given the weightage (n) of 0.75 and $R_{duration}$ is given the weightage $(1-n)$ of 0.25. This result is consistent across all the recommendation methods. Thus we can conclude that assigning weightage to both frequency and duration helped us in achieving more accurate recommendations.

6 Conclusion and Future Work

In this paper, we have presented an approach to tackle the problem of recommending shows on linear TV by converting the implicit feedback from the users (collected in terms of frequency and duration of user-show interactions) to explicit ratings. For each unique user-show interaction we derive two ratings, frequency rating and duration rating and then the two ratings are combined together to obtain a final rating. Experimentally we have verified that recommendations generated are more accurate when both frequency and duration ratings are given some weightage to compute final rating than when only one of them are considered separately. The focus of our future work is to make the final rating calculation more accurate by assigning optimum weightage to frequency and duration ratings using some machine learning technique.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
2. Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Difino, A., Negro, B.: User modeling and recommendation techniques for personalized electronic program guides. In: Personalized Digital Television. Human-Computer Interaction Series, vol. 6, pp. 3–26 (2004)
3. Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98). pp. 43–52 (1998)
4. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized non-negative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1548–1560 (2011)
5. Chang, N., Irvan, M., Terano, T.: A tv program recommender framework. Procedia Computer Science 22, 561–570 (2013)
6. Cotter, P., Smyth, B.: Ptv: Intelligent personalised tv guides. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. pp. 957–964 (2000)
7. Cremonesi, P., Modica, P., Pagano, R., Rabosio, E., Tanca, L.: Personalized and context-aware tv program recommendations based on implicit feedback. In: Stuckenschmidt H., Jannach D. (eds) E-Commerce and Web Technologies. LNBIP. vol. 239 (2015)
8. Das, D., Horst, H.: Recommender systems for tv. In: Workshop on Recommender Systems, Proceedings of 15th AAAI Conference, pp. 35–36 (1998)
9. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
10. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08). pp. 263–271 (2008)
11. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer Society 42(8), 30–37 (2009)
12. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. Advances in Neural Information Processing Systems 20, 1257–1264 (2008)
13. Su, X., Khoshgoftaar, T.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence 2009 (2009)
14. Turrin, R., Condorelli, A., Cremonesi, P., Pagano, R.: Time-based tv programs prediction. In: RecSysTV Workshop at ACM RecSys 2014, pp. 957–964 (2014)

Study of Adaptive Model Predictive Control for Cyber-Physical Home Systems

Sian En OOI[1], Yuan FANG[1,2], Yuto LIM[1], and Yasuo TAN[1]

[1] Japan Advanced Institute of Science and Technology (JAIST),
1-1 Asahidai, Nomi, Ishikawa, 923-1292 JAPAN,
{sianen.ooi, yfang, ylim, ytan}@jaist.ac.jp,
WWW home page: http://www.jaist.ac.jp
[2] Dalian Polytechnic University (DPU),
No.1 Qinggongyuan, Dalian, Liaoning, CHINA

Abstract. With the inception of connected devices in smart homes, the need for user adaptive and context-aware systems have been increasing steadily. In this paper, we present an adaptive model predictive control (MPC) based controller for cyber-physical home systems (CPHS) environment. The adaptive MPC controller is integrated into the existing Energy Efficient Thermal Comfort Control (EETCC) system that was developed specifically for the experimental smart house, iHouse. The proposed adaptive MPC is designed in a real time manner for temperature reference tracking scenario where it is evaluated and verified in a CPHS simulation using raw environmental data from the iHouse.

Keywords: adaptive, model predictive control, smart homes, cyber-physical systems

1 Introduction

Recent growth in home automation research affirms the importance on enhancing the quality of life (QoL) in residential and commercial buildings [1–6]. Home automation typically requires key elements such as sensing, actuation and control. These key elements forms the cores of cyber-physical systems (CPS), which justifies its place in smart home environments. One of the active research in smart homes domain are energy efficient thermal comfort, where building architecture, envelop, heating, ventilation and air conditioning (HVAC) and control are within its scope. Model based controls such as model predictive control (MPC) have gained traction throughout the years especially in applications such as thermal comfort control [1, 7, 6]. Some of the advantages of MPC in thermal comfort control application are its capability to apply anticipated control strategies in lieu of corrective strategies while simultaneously handling multiple objectives and constraints. However, model based control normally require expert knowledge of the entire process to design and tune the plant model to accurately represent the actual control plant. Practical implementations of model based control for smart homes are generally unrealistic as every room or building have different thermal and insulation characteristics.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_17

In this paper, the objective is to address presents an adaptive MPC controller for cyber-physical home systems (CPHS) environment. The main goal for this paper is twofold: (i) to implement adaptive MPC based temperature controller for CPHS; and (ii) to implement real time control based on CPS approach. With adaptive model based control, model tuning effort should be reduced significantly as it automatically identify the plant characteristics and tune the controller parameters at runtime. The rest of the paper is organized as follows. Section 2 introduces the background on relevant topics to this paper. The experimental house and its system, adaptive MPC controller and online model estimator details are described in Section 3. Proposed controllers are simulated during autumn season while its results and discussions are presented in Section 4. Finally, some relevant conclusions are summarized in Section 5.

2 Research Background

2.1 Cyber-Physical Home Systems

CPS are described as systems where their physical and computational elements are strictly interlinked together by networking elements [4]. This mechanism is incorporated into smart home environment to form CPHS, where it is comprised of the physical and cyber worlds interlinked together by various communication networks. Sensing and actuating domain are part of the physical world in a CPHS environment while the computing elements such as data storage and supervisory control are part the control domain in the cyber world. One of the implementation of smart homes are the iHouse, which is an advanced experimental smart house, located at Nomi City, Ishikawa prefecture, Japan. It is a conventional two-floor Japanese-styled house featuring more than 300 sensors, home appliances, and electronic house devices that are connected using ECHONET Lite version 1.1 and ECHONET version 3.6 [4]. The EETCC system designed in previous work was based on the CPS approach, where its implementation in the iHouse can be found in [4]. The EETCC system tightly coupled appropriate sensors and actuators together while a state based supervisory controller performs relevant control to maintain the thermal comfort level in a room. The state based supervisory controller is a rule based algorithm those objective is to promote energy efficiency by prioritizing the use of natural resources to maintain the thermal comfort level in a room rather than the use of HVAC. However, this supervisory controller suffers from non-optimal control strategies as it senses the changes in thermal comfort level without anticipating any future events.

3 Adaptive MPC for EETCC System

The control plant in this paper is based on the iHouse, where various types of networked sensors and actuators are linked together to provide the necessary feedback parameters and output controls to the proposed controller. The EETCC system introduced in [4] is used as the CPHS platform, where its architecture is illustrated in Fig. 1. The EETCC system is comprised of three main components: (i) controller; (ii) network and communication; and (iii) plant.

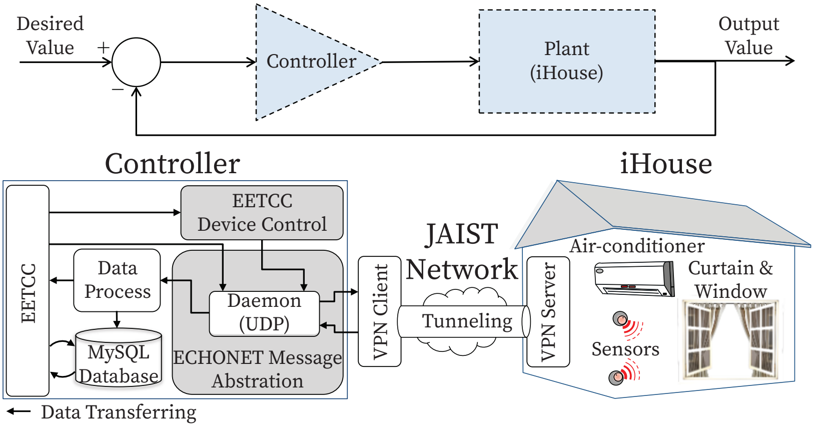


Fig. 1. Architecture of EETCC system.

In this paper, an adaptive MPC is employed as the local level control of the HVAC while the EETCC algorithm introduced in [4] is employed as supervisory level control in the iHouse EETCC system. Proportional Integral (PI) control algorithm is commonly implemented as local level control due to its low requirement for computing resource while supervisory level control are commonly governed by rule-based control (RBC) and optimal control algorithms such as MPC that are resource intensive. An example of such setup can be found in [1]. Advancement in embedded computing allows MPC to be used in local level control that are real time in general.

3.1 Adaptive Model Predictive Control

In this section, an adaptive MPC with online model estimation system is introduced. The plant in this context is the iHouse bedroom, subjected to outdoor environment disturbances as well as HVAC as its input. This is illustrated in Fig. 2. The MPC controller block is comprised of MPC internal plant model and state estimator as its prediction block, and the optimization block that computes input optimization with respect to the imposed cost and constraints.

The thermodynamic characteristics of the plant modeled using heat equations are comprehensively explained in [6]. Since the controller is a discrete MPC, the plant is transformed into a thermal resistor-capacitor (RC) model those form is a discrete state-space model as shown in [6]. Hence, a discrete state-space model can be given by the following equations

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + Wv(k) \\ y(k) &= Cx(k) + v(k) \end{aligned} \tag{1}$$

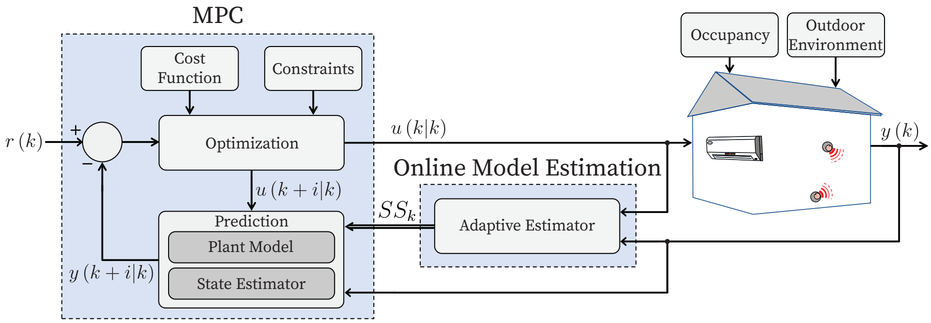


Fig. 2. Adaptive MPC system model.

where x is the state vector, u is the input vector, y is the output vector while A, B, C and W are constant state-space matrices of coefficients. $v(k)$ is the disturbance vector at interval k that is consisted of heat gain from outdoor temperature and solar radiation. The MPC internal plant model in this paper is a simplified plant, which only considers the HVAC input and room temperature output. Outdoor environmental disturbances are excluded from MPC prediction.

The MPC controller and online model estimation block are both in discrete time while the plant is in continuous time. Zero hold order discretization is employed for the inputs and outputs of the EETCC system controller, where the signals are held constant during the sampling period until the next sampling instance. This introduced delay into the system which deteriorates the closed loop performance of non-adaptive MPC controller. Hence, the working of how adaptive MPC with online state estimation can reduce internal model discrepancy as well as managing input delay issues are explained in the following subsections.

State Estimation Besides the internal plant model, state estimation is also one of the components in MPC prediction block. This section briefly describe a general state estimator in a MPC controller, Kalman filter (KF) as well as the linear time varying Kalman filter (LTVKF) state estimator in adaptive MPC. The state estimate of a MPC controller in the MPC toolbox provided by Simulink is described in [8], where the controller internal state estimation at interval k and the updated state estimation at interval $k+1$ are given by

$$x_c(k \mid k) = x_c^{rev}(k \mid k-1) + M_k e(k) \tag{2}$$

$$x_c(k+1 \mid k) = A_c x_c^{rev}(k \mid k-1) + B_u u^{opt}(k) + B_v v(k) + L_k e(k) \tag{3}$$

where $x_c^{rev}(k \mid k-1)$ is the predicted state estimate at interval k, $e(k)$ is the difference between measured state and predicted state estimate at interval k, $u^{opt}(k)$ is the optimal input computed by MPC at interval k, $v(k)$ is the input disturbance at interval k, A_c is the internal plant model state-space matrices of coefficients, B_u and B_v are the internal plant state vector for input control and disturbance, L_k and M_k are the KF gain matrices at interval k. Conventional

MPC pre-calculate the KF gain matrices during initialization of the MPC controller and keep the gain matrices constant throughout the entire runtime of the controller. This relies on the accuracy of the internal plant model during design stage, which requires expert knowledge of the entire control process and tedious parameter tuning to accurately represent the actual control plant. However, this shortcomings can be alleviated by employing adaptive MPC. Basically, adaptive MPC works by updating of the KF gain matrices at every interval of k, which calibrates the controller internal state according to the latest measurement. The KF gain matrices at interval k are given by

$$L_k = \left(A_k P_{k|k-1} C_{m,k}^T + N\right)\left(C_{m,k} P_{k|k-1} C_{m,k}^T + R\right)^{-1} \tag{4}$$

$$M_k = P_{k|k-1} C_{m,k}^T \left(C_{m,k} P_{k|k-1} C_{m,k}^T + R\right)^{-1} \tag{5}$$

where $P_{k|k-1}$ is the state estimate error covariance matrix at interval k based on the measured state at interval $k-1$, A_k and $C_{m,k}$ are the state-space matrices of coefficients updated at interval k, N and R are the state estimation constant covariance matrices [8].

Optimization Problem One of the advantages of MPC is its ability to handle multiple objectives and constraints. MPC generally solves a quadratic problem (QP) at each interval to find the optimal control input with respect to the objectives and constraints. This paper implements a temperature reference tracking MPC controller bounded by the HVAC capability and plant temperature boundary. A general objective or cost function that penalizes reference signal deviation, large change in control input and constraint violations can be given by

$$\begin{aligned} J_k = &\sum_{i=1}^{n_y} \left\{w_i^y \left[r\left(k+1|k\right) - y\left(k+1|k\right)\right]\right\}^2 \\ &+ \sum_{i=0}^{n_u-1} \left\{w_i^{\Delta u}\left[\Delta u\left(k+1|k\right)\right]\right\}^2 + \rho_e \varepsilon_k^2 \\ \text{s.t.}& \\ &y_{min} \le y\left(k+1|k\right) \le y_{max} \\ &u_{min} \le u\left(k+1|k\right) \le u_{max} \end{aligned} \tag{6}$$

where n_y is the controller prediction horizon, n_u is the input control horizon, w_i^y is the plant output tuning gain at ith prediction step, $w_i^{\Delta u}$ is the change in input control tuning gain at ith prediction step, r is the reference signal, y is the predicted plant output and $\Delta u\left(k+1|k\right)$ is the change in the optimal input control at time $k+i$ computed at interval k, ρ_e is the constraint violation penalty gain and ε_k is the constraint slack variable at interval k. The HVAC minimum and maximum saturation points are given by u_{min} and u_{max} respectively while the plant minimum and maximum room temperature are given by y_{min} and y_{max}.

3.2 Online Model Estimation

The workings of adaptive MPC are described in the previous section, where it relies on model parameters updates at every interval to tune its internal state estimator. This section briefly describes the online model estimator that is used in this paper to update the adaptive MPC internal state estimator. Several methods are employed in literature related to online model estimation in buildings, such as extended KF (EKF) in [2], unscented KF (UKF) in [2, 9, 3] and controlled autoregressive integrated moving average (CARIMA) in [10]. Besides, previous work on the iHouse involves an autoregressive moving average (ARMA) model based offline estimation in [5]. This paper takes a different approach on model estimation of the iHouse, where online estimation is employed and input from HVAC is also taken into account. An autoregressive moving average with exogenous input (ARMAX) model is used in this paper to represent the iHouse. Thus, the plant dynamic thermal behavior can be expressed in a black box ARMAX model as

$$A(q)y(t) = B(q)u(t - nk) + C(q)e(t) \tag{7}$$

where y is the plant room temperature, u is HVAC control input, e is the noise term with zero mean. The parameters of autoregressive (AR) are given by $A(q) = 1 + a_1 q^{-1} + \cdots + a_{na} q^{-na}$ and $a_1, \cdots, a_{na}$, where na is the AR order. The parameters of exogeneous (X) input are given by $B(q) = b_1 + b_2 q^{-1} + \cdots + b_{nb} q^{-nb+1}$ and $b_1, \cdots, b_{nb}$, where nb is the X input order. The parameters of moving average (MA) are given by $C(q) = 1 + c_1 q^{-1} + \cdots + c_{nc} q^{-nc}$ and $c_1, \cdots, c_{nc}$, where nc is the MA order. Since the MPC internal model in this paper is intended to be a simplified plant model, the na, nb and nc are configured as first order to reduce its complexity.

Two estimation algorithms are employed to identify the ARMAX model parameters: (i) normalized least mean square (LMS) and (ii) KF algorithm. This two parameter estimation algorithms can be generally described by

$$\hat{\theta}(t) = \hat{\theta}(t - 1) + K(t)\left[y(t) - \hat{y}(t)\right] \tag{8}$$

where $\hat{\theta}$ is the estimated ARMAX model parameter, K is the prediction error gain, y is the measured output and $\hat{y}$ is the predicted output. In-depth details of the two algorithms can be found in [11]. Although the ARMAX model parameters can be pre-estimated if the initial conditions are known, an initial guess is used instead during the initialization process to represent a more realistic deployment of such system. Besides, KF algorithm is capable of managing parameter estimation uncertainty during initialization by tuning the initial parameter covariance. Typically, a large initial parameter covariance is used when high uncertainty exists in the initial estimation. Uncertainty in the state equations and process noises can also be managed by tuning KF parameter covariance. While Simulink does not allow online tuning of the delay parameter, the expected delay value can be pre-configured in the online model estimator. Furthermore, the ARMAX model is converted into an equivalent state-space form before propagating the updated parameters to the adaptive MPC controller.

4 Numerical Evaluation

This section examines the performance of the proposed adaptive MPC controller with online model estimation in a CPHS environment, where a conventional MPC controller is employed as baseline controller. The term “MPC” and “conventional MPC” are used indistinguishably in this section. The simulation is built based on an actual bedroom in the iHouse and simulated on Simulink R2017b. Environmental sensors are polled and stored in the EETCC database at an interval of 10 seconds. Hence, the zero hold order sampling interval is also set to 10 seconds while MPC prediction and control horizon are configured to 2 minutes. Besides, the simulation outdoor environment is based on measured data from the iHouse on 1st November 2013 as shown in Fig. 3 while the simulation is performed on a MacBook Pro with Intel Core i7 processor at 3.1 GHz and 16 GB. The remaining simulation parameters are listed in Table 1.

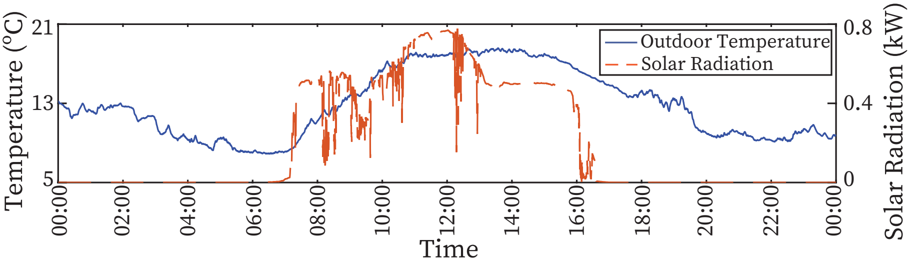


Fig. 3. Outdoor environment on 1st November 2013.

Table 1. Simulation parameters and settings.

Parameter	Value
Volume of room $(L \times W \times H)$, V_{room}	$5.005 \times 4.095 \times 2.4\,\mathrm{m}^3$
Density of air	$1.2\,\mathrm{kg/m}^3$
Specific heat capacity of air	$1.005\,\mathrm{kJ/kg^\circ C}$
Air volume flow rate, CFM	$300\,\mathrm{ft}^3/\mathrm{min}$
Minimum cooling load of HVAC, u_{min}	5 kW
Maximum cooling load of HVAC, u_{max}	6.3 kW
Coefficient of performance, COP	3.44
Area of window type 1, A_{w1}	$1.815\,\mathrm{m}^2$
Area of window type 2, A_{w2}	$0.66\,\mathrm{m}^2$
U-value of window type 1, u_{w1}	$3.4\,\mathrm{W/m^2{}^\circ C}$
U-value of window type 1, u_{w2}	$1.7\,\mathrm{W/m^2{}^\circ C}$
Solar transmittance of window type 1, g_{w1}	0.79
Solar transmittance of window type 2, g_{w2}	0.41

4.1 Results and Discussion

Both MPC and adaptive MPC controllers are subjected to the same cost function and constraints as shown in Eqn. (6). Since the cost function is a reference tracking, a step signal of 25 °C is applied as the reference at 00:00 to the end of the simulation. Besides, normalized LMS and KF algorithm in the online model estimator are both evaluated as part of the adaptive MPC controller. Responses of all MPC controllers and online model estimators during initialization are shown in Figure 4. MPC reached steady state faster than both adaptive MPC controllers as the online model estimators requires a number of iterations before achieving stabilization. The online model estimator system matrix, A converged from its initial value to a stable value during the first two minutes of initialization process. During this period, the adaptive MPC applied large rate of change for the input to reach both ends of the HVAC constraints to quickly identify and tune the ARMAX model to the characteristics of the plant. Updated state space parameters are propagated to the adaptive MPC at the same time, where HVAC control inputs are promptly corrected from maximum cooling mode to heating mode as observed in Figure 4. Similar input response is also observed in [10].

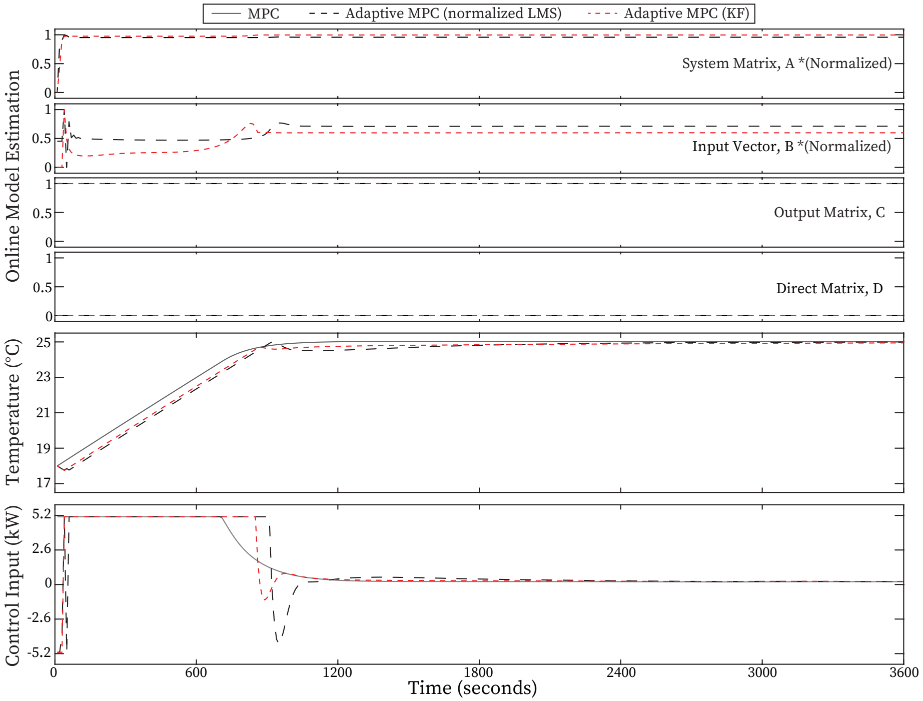


Fig. 4. Initial responses of MPC and adaptive MPC for iHouse bedroom. Note: system matrix and input vector are normalized.

Since the performance of adaptive MPC is dependent on the online model estimator, adaptive MPC with KF based online model estimator performed better as it offered faster convergences, less oscillations and overshoots than normalized LMS algorithm as shown in Figure 4. Besides, the steady state responses of all MPC controllers and online model estimators when subjected to disturbance from outdoor environment are shown in Figure 5. Unlike adaptive MPC, MPC is susceptible to random disturbances as it is unable to correct its internal model errors. The upper output error of MPC is 0.173 °C while the lower output error is 0.168 °C. The upper output errors for adaptive MPC with normalized LMS and KF based online model estimators are 0.0246 °C and 0.0141 °C while the lower output errors are 0.0208 °C and 0.0183 °C respectively. Although KF performed better than normalized LMS, KF is computationally heavier. However, computational load from various online model estimator algorithms are insignificant as it only constitute not more than 1.5% of the total simulation time while MPC optimization process constitute more than 86.5%. Average simulation time for MPC and adaptive MPC with online model estimation is 43.6 and 291.3 seconds, where the difference is more than six times. Thus, trade-offs between performance and computational cost should be assessed for practical implementations.

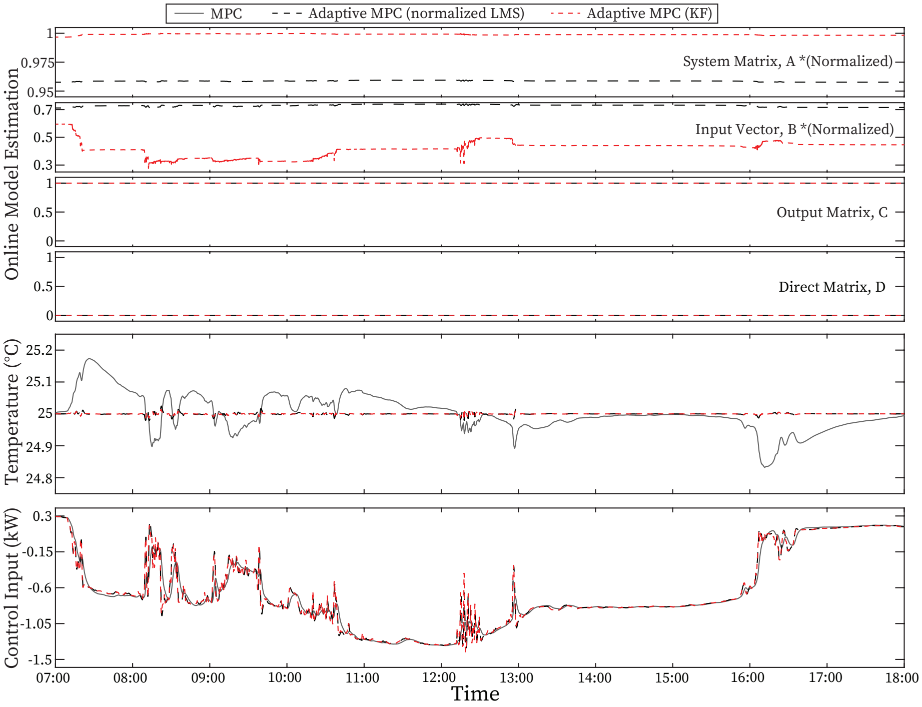


Fig. 5. Steady state responses of MPC and adaptive MPC with disturbance for iHouse bedroom. Note: system matrix and input vector are normalized.

5 Concluding Remarks

This paper summarizes the implementation details of real time adaptive MPC with online model estimation in CPHS environment. The adaptive MPC with various online model estimators are simulated and benchmarked against a conventional MPC controller to evaluate the performance and advantage of adaptive capability in a CPHS environment. Both adaptive MPC with KF and normalized LMS based online model estimators showed improvements in temperature reference tracking scenarios. However, computation required will increase if self adaptive capability and strict reference regulation are required. System designers should take note of the trade-offs between performance and computation cost for practical implementations of predictive control in a CPHS environment.

References

1. F. Barata, J. Igreja, and R. Neves-Silva, "Model predictive control for thermal house comfort with limited energy resources," in *CONTROLO2012*, 2012.
2. M. Maasoumy, B. Moridian, M. Razmara, M. Shahbakhti, and A. Sangiovanni-Vincentelli, "Online simultaneous state estimation and parameter adaptation for building predictive control," in *ASME 2013 Dynamic Systems and Control Conference*, pp. V002T23A006–V002T23A006, American Society of Mechanical Engineers, 2013.
3. P. Radecki and B. Hencey, "Online model estimation for predictive thermal control of buildings," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 4, pp. 1414–1422, 2017.
4. Y. Lim, S. E. Ooi, Y. Makino, T. K. Teo, A. Rayner, and Y. Tan, "Implementation of energy efficient thermal comfort control for cyber-physical home systems," *Advanced Science Letters*, vol. 23, no. 11, pp. 11530–11534, 2017.
5. S. F. M. Hussein, M. A. A. Bakar, Y. Makino, H. Nguyen, S. S. Abdullah, Y. Lim, and Y. Tan, "Simplifying the auto regressive and moving average (arma) model representing the dynamic thermal behaviour of ihouse based on theoretical knowledge," in *Modeling, Design and Simulation of Systems*, (Singapore), pp. 697–711, Springer Singapore, 2017.
6. S. E. Ooi, Y. Makino, Y. Fang, Y. Lim, and Y. Tan, "Study of predictive thermal comfort control for cyber-physical smart home system," *IEICE Technical Report*, vol. 117, no. 426, pp. 29–34, 2018.
7. A. Afram and F. Janabi-Sharifi, "Theory and applications of hvac control systems– a review of model predictive control (mpc)," *Building and Environment*, vol. 72, pp. 343–355, 2014.
8. A. Bemporad, M. Morari, and N. L. Ricker, "Model predictive control toolbox 3 user's guide," *The MathWorks Inc.*, 2010.
9. A. Martinčević, A. Starčić, and M. Vašak, "Parameter estimation for low-order models of complex buildings," in *Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2014 IEEE PES*, pp. 1–6, IEEE, 2014.
10. M. Short and F. Abugchem, "A microcontroller-based adaptive model predictive control platform for process control applications," *Electronics*, vol. 6, no. 4, p. 88, 2017.
11. L. Ljung, "System identification toolbox," *The MathWorks Inc.*, 1988.

Implementation of Constraint Programming and Simulated Annealing for Examination Timetabling Problem

Tan Li June, Joe H. Obit, Yu-Beng Leau and Jetol Bolongkikit

Knowledge Technology Research Unit, Universiti Malaysia Sabah,
Labuan International Campus, Malaysia
june_tan93@hotmail.com, joehenryobit@gmail.com,lybeng@ums.edu.my
jetol@ums.edu.my

Abstract. Examination timetabling problems is the allocation of exams into feasible slots and rooms subject to a set of constraints. Constraints can be categorized into hard and soft constraints where hard constraints must be satisfied while soft constraints are not necessarily to satisfy but be minimized as much as possible in order to produce a good solution. Generally, UMSLIC produces exam timetable without considering soft constraints. Therefore, this paper proposes the application of two algorithms which are Constraint Programming and Simulated Annealing to produce a better solution. Constraint Programming is used to generate feasible solution while Simulated Annealing is applied to improve the quality of solution. Experiments have been conducted with two datasets and the results show that the proposed algorithm managed to improve the solution regardless the different problem instances.

Keywords: Examination timetabling, Constraint Programming, Simulated Annealing.

1 Introduction

Examination timetabling problem is one of the most concerns under domain of academic institution [17]. The problem is to assign events into timeslots and rooms by satisfying a set of constraints. Constraints are categorized as hard constraints and soft constraints. Hard constraints are used to determine the feasibility of the timetable. Thus, all hard constraints must be satisfied in any circumstances. For example, no student can sit for more than one exam simultaneously. While soft constraints are used to determine the quality of the timetable, however, soft constraints are not necessarily to be all satisfied but should be minimized as much as possible in order to produce a better quality timetable. For example, minimize the number of students with consecutive exams [22]. Examination timetabling problems have been widely studied with different approaches [3, 5, 9, 15, 17]. This research developed a model based on the datasets from Universiti Malaysia Sabah Labuan International Campus (UMSLIC) with a set of hard constraints and soft constraints. Technically, this research involves

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_18

two phases (1) Initialization phase: to generate feasible timetable by using constraint programming (2) Improvement phase: to enhance the quality of timetable with Simulated Annealing. In initialization phase, all hard constraints must be satisfied to generate feasible timetable as an initial solution. After generating the initial solution, next phase is to apply simulated annealing to improve it by reducing the soft constraints as much as possible.

The paper is organized as follows. After the introduction of the research in Section 1, Section 2 describes previous work regarding to the same problem domain. Section 3 provides explanation on the problem and the set of constraints of the problem are presented in mathematical formulas in Section 4. Section 5 explains the implementation of algorithms on the problem and the experimental results are discussed in Section 6. Section 7 concludes this research.

2 Related Works

Timetabling problems have been widely studied since 1960's with the implementation of different approaches [13, 19].The complexity of the problem make it more difficulty to produce an optimal solution or near to optimality. It is known that every institution has different policies and requirements therefore, not every approach applied in the literature can perform well in other problems event in the same problem domain dataset. Currently, UMSLIC uses CELCAT system to generate examination timetable. The timetable is feasible however, the system does not take any soft constraints into consideration which causes the low quality of timetable. For example, many students with consecutive exams in a day which causes students have limited time to do revision. As stated previously, this research involves two phases. The first phase implements constraint programming to produce feasible timetable and it will be further improved with the application of simulated annealing in the second phase. In [8], the course timetabling problem of UMSLIC had been studied and constraint programming algorithm was developed to solve all the hard constraints. The result shows the constraint programming is able to produce feasible timetable within a short period of time with several times of experiments.

Meanwhile, [20] studied on school timetabling problem by integrating Constraint Programing with operations research produced results are close to pre-defined optimal values. Overall, the solutions were produced in an acceptable duration. In [7], they implemented four different cooling strategies of Simulated Annealing to improve the quality of course timetable for every semester. They are linear cooling, exponential cooling, linear multiplicative cooling and geometric cooling. Among these four strategies, geometric cooling performed better than the rest which showed the best improvement from initial solution.

Research [19] proposed the combination of constraint programming and Simulated Annealing to solve examination timetabling problem of HoChiMinh City (HCMC) University of Technology. This research used Kempe chain to determine starting temperature which could improve the performance of Simulated Annealing in terms of its efficiency.

Besides, [12] implemented four hybrid heuristics to generate feasible solution for University course timetabling problem which are Sequential Heuristics (Largest Degree and Saturation Degree), Local Search, Tabu Search and Great Deluge. The results showed that all four hybrid heuristics could generate feasible solution but none of them perform outstandingly in terms of quality of solution such as this heuristic cannot scheduling students to have less than two consecutive course in a day. However, result from [12] shows that Sequential Heuristics could only produce feasible solution for small instances of the Socha et al [18] dataset.

In [14], Non-linear Great Deluge (NLGD) was proposed to compare its performance with the conventional Great Deluge (GD). NLGD algorithm modifies the conventional GD in the way of changing in water level with non-linear decay rate. The modification shows that NLGD outperforms over conventional GD and other algorithms in terms of the scheduling computational time and the improvement of the solutions. Research from [11] proposed the Evolutionary Non-Linear Great Deluge (ENGD) for course timetabling problem. [11] extends NLGD with effective operators and three neighborhood moves for solving the instances of Socha et al [18]. The results showed that ENGD perform effectively, obtaining best solution with zero penalties in small instances.

3 Problem Background

Examination timetabling problem is basically formed by four sets of parameter: exams, rooms, timeslots and constraints. This research aims to study examination timetabling problem of UMSLIC and schedule exams into room and timeslots subject to a set of constraints. Constraints are obtained from the Academic Service Division (BPA) of UMSLIC through several times of interview. Constraints are categorized into hard and soft constraints and shown as below:

Table 1. Hard constraints and soft constraints

Hard Constraints	Soft Constraints
HC_1: All exams must be assigned into available timeslots.	SC_1: Maximize the room utilization.
HC_2: No student can attend more than one exams simultaneously.	SC_2: Minimize the number of students with consecutive exams.
HC_3: The room capacity must be equal or greater than the total number of students taking the particular exams.	SC_3: Prioritize the exams with greater size in first two weeks.

UMSLIC uses CELCAT system to generate examination timetable which do not consider the soft constraints of the problem, hence, the timetable is feasible but low quality. Therefore, this research applies simulated annealing to further improve the solution quality by reducing the soft constraints violations as much as possible. In terms of resources, the duration of whole examination is three weeks, with 11 slots per week. Therefore, 33 slots in total and six rooms are available to accommodate the exams. However, some exams have large student enrollment which cannot be accommodated into the largest examination hall of UMSLIC. Therefore, it is necessary to assign those exams into multiple rooms. This research carried out experiments with two different semester datasets which are semester 2 session 2014/2015 and semester 1 session 2015/2016. Both datasets have different total number of students and exams. Table 2 summarizes the attributes of both datasets respectively.

Table 2. Summary of datasets

	Semester 2 session 2014/2015	**Semester 1 session 2015/2016**
Students	2248	2371
Exams	112	125
Timeslots	33	33
Rooms	6	6

4 Formulation Model

In order to achieve the objective of this research, formulation model has been developed to evaluate the feasibility and quality of the timetable. The formal model of examination is presented as below:

- *$E = E_1, \ldots, E_n$ where n is the total number of examinations*
- *$T = T_1, \ldots, T_t$ where t is the total number of timeslots*
- *$R = R_1, \ldots, R_r$ where r is the total number of rooms*
- *$C = C_1, \ldots, C_c$ where c is the total number of room combinations*
- *$S = S_{E1,} \ldots, S_{En}$ is the list of total number of students taking the exam E where n is the total number of examinations*

In this research, the feasibility and quality of the solution is evaluated based on the constraints violations. There are three hard constraints and three soft constraints in this research and each of them are assigned with different penalty cost. Table 3 shows the penalty cost of each constraint. For example, if a student is scheduled to attend more than one exam at the same time, the solution is violated with second hard constraints. Hence, the penalty cost of the solution is 100,000.

Table 3. Penalty cost of constraints violations

Weight	Penalty	Description
λ_1	100,000	Hard constraints, HC_1
λ_2	100,000	Hard constraints, HC_2
λ_3	100,000	Hard constraints, HC_3
λ_4	1	Soft constraints, SC_1
λ_5	1	Soft constraints, SC_2
λ_6	1	Soft constraints, SC_3

The penalty cost of the solution, F is presented in equation (1)

$$F = SC_1 + SC_2 + SC_3 \tag{1}$$

subject to the total of HC_1, HC_2, HC_3 must be zero, as stated previously, all hard constraints must be satisfied in order to produce a feasible solution. HC_1 is to ensure all exams are assigned and formulated as equation (2)

$$HC_1 = \lambda_1 \sum_{i=0}^{n} E_i \tag{2}$$

where n is the total number of exams, λ_1 is HC_1's weight, E_i is the number of exams that cannot be assigned. HC_2 is to prevent clashes happen in timetable and formulated as equation (3)

$$HC_2 = \lambda_2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} EE_{ij} \tag{3}$$

where n is the total number of exams, λ_2 is HC_2's weight, EE_{ij} is when exam i and j clash with each other. HC_3 is to assign exams into rooms based on the size of the exams and formulated as equation (4)

$$HC_3 = \lambda_3 \sum_{i=1}^{n} \sum_{j=1}^{r} ER_{ij} \tag{4}$$

where n is the total number of exams, r is the total number of rooms, λ_3 is HC_3's weight, ER_{ij} is the number when the size of exam i is larger than capacity of room j. There are three soft constraints in this problem, SC_1 is to reduce the extra space of the room and formulated as equation (5)

$$SC_1 = \lambda_4 \sum_{i=1}^{n} \sum_{j=1}^{r} ER_{ij} \tag{5}$$

where n is the total number of exams, r is the room size of the exam, λ_4 is SC_1's weight, ER_{ij} is the extra room space of the assignment of exam i into room j.

SC_2 is students should not sit for more than one subject a day and formulated as equation (6)

$$SC_2 = \lambda_5 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij} \tag{6}$$

where n is the total number of exams, λ_4 is SC_2's weight, E_{ij} is the time interval between exam i and exam j. SC_3 is to assign large exam into first two weeks with the evaluation of equation (8), in order to get large exam, equation (7) will be used to identify which exam is considered as a large exam

$$average_s = \frac{S}{n} \tag{7}$$

$$SC_3 = \lambda_6 \sum_{i=\frac{T}{2}}^{t} \sum_{j}^{n} TE_{ij} \tag{8}$$

where S is the total size of all exams, n is the total number of exams, t is the total number of timeslots, λ_5 is SC_3's weight, TE_{ij} is the number when exam i has larger value than $average_s$ and it is assigned later than first two weeks.

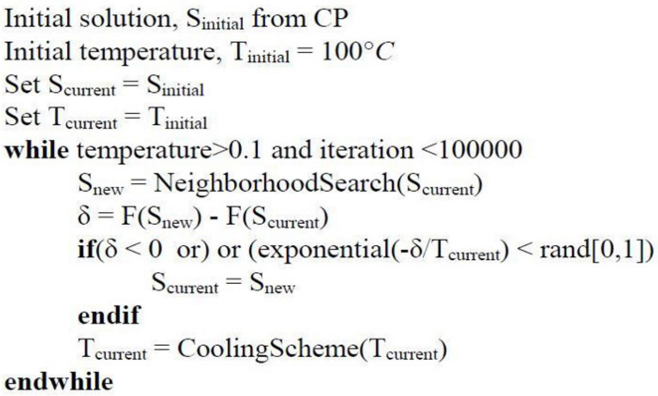
5 Implementation

The research work proposed in this paper approaches examination timetabling problem in two stages. For the first stage, Constraint Programming (CP) is applied to generate feasible solution by solving all the hard constraints and it is suitable for timetabling problem as it performs effectively when constraints of the problem are presented into numeric form [6]. For example, in [8], constraints are presented in binary matrices such as course conflict matrix to identify the conflict between courses. Research [20] implemented CP to solve the school timetabling problems which produce very good result.

In this research, all the exams are categorized into 4 parts: main exams, Promotion of Knowledge and Language (PPIB) exams, language exams and co-curriculum exams. Categorizing exams into different types is due to the reason that some exams have to be assigned into specific timeslot. For example, co-curriculum exams have to be on Saturday. Besides, a list of room combinations (RC) is also created for those exams have too many students which cannot be accommodated into the largest examination hall of UMSLIC. Therefore, room sharing is necessary for these exams.

At the beginning of the algorithm, a pool of unscheduled exams (UE) is generated to store all exams. The sequence of exams assignment is co-curriculum exams, language exams and others. CP will firstly select an exam according to the sequence of exams from UE and a timeslot at random. If the timeslot is feasible for the exam, it will select a room from RC at random. If the room is available and large enough to fit the exam, the exam will be assigned into that feasible slot and room. Exam will then moves from UE to pool of scheduled exams (SE) which indicates that exam is assigned successfully. If no room is available in that particular slot, the algorithm will select another slot until the exam is assigned. The whole step will iterate until UE is empty. However, it could happen when some exams are not able to fit into any slot due to limited room choice or clashes and those unassigned exams will move to pool of fail scheduled exams (FE) in order to empty UE and identify which exam is fail to assigned. Meanwhile, if UE is empty and there are still some exams in FE, neighbourhood search will take place in this situation to disturb the solution. An exam from SE, a slot and a room will be selected randomly to perform swapping. If the slot and room is feasible, the exam will move that slot and room. Hence, the solution is changed and proceeded with the scheduling of exams from FE. The search will stop until FE is empty.

In the second stage, Simulated Annealing (SA) is applied once the feasible solution is found to improve the timetable quality. SA algorithm is derived from the idea of cooling process in a physical system to study on the movement of the particles. The process of SA in timetabling problem can be represented with the elements of SA [4]. The physical system refers to the pool of solution and it is made up by smaller particles which refer to each solution. Meanwhile, the penalty cost of the timetable is represented by the system energy. In this research, the penalty cost of initial solution and initial temperature are defined. In cooling process of SA, the searching area is getting smaller as the temperature drops and the movement of particles becomes less active.
The result of [1] shows that SA performs outstandingly with the implementation of adaptive cooling and reheating scheme in SA for solving course timetabling problem of Syracuse University. In this research, SA with geometric cooling scheme is applied in second stage to improve the solution created by CP. The pseudocode of SA is shown in Fig.1:



```
Initial solution, $S_{initial}$ from CP
Initial temperature, $T_{initial} = 100°C$
Set $S_{current} = S_{initial}$
Set $T_{current} = T_{initial}$
while temperature>0.1 and iteration <100000
    $S_{new}$ = NeighborhoodSearch($S_{current}$)
    $\delta = F(S_{new}) - F(S_{current})$
    if($\delta < 0$  or) or (exponential($-\delta/T_{current}$) < rand[0,1])
        $S_{current} = S_{new}$
    endif
    $T_{current}$ = CoolingScheme($T_{current}$)
endwhile
```

Fig. 1 Pseudocode of SA

In this phase, the initial temperature of SA in the experiments is set at 100°C. After CP created a solution, it will be used as an initial solution ($S_{current}$). As the temperature decreases, the algorithm will search for a new solution (S_{new}). If the penalty cost of S_{new} is lower than $S_{current}$, $S_{current}$ will be updated as S_{new}. In order to ensure SA performs effectively, different numbers of iterations were tested: 1000, 10000 and 100000. 100000 iterations performs best among them. Therefore, the process will iterate until it reaches 100000 iterations or the temperature drops to 0°C. The geometric cooling scheme is presented with equation (9):

$$t = \alpha t \tag{9}$$

where t is the temperature, α is the reduction parameter for geometric cooling, this research set it at 0.9999 according to [4].

When the temperature drops, the search space will decrease which could lead to the solution stuck in local optima [2]. Therefore, there is an acceptance criteria [1] known as Boltzmann probability, P with equation (10) which will accept worse solution with a certain probability.

$$P = e^{-\delta/t} \tag{10}$$

where δ is the difference of the penalty cost between current solution and new solution.

6 Experimental Results

This research is carried out with two different semester datasets of real world from UMSLIC as described in Table 2. Both datasets have different number of students and exams. Therefore, the research produced different result although they are solved by same algorithm.

The experiments in the initialization phase of this research which is carried out by using CP to solve all the hard constraints to produce feasible solutions. CP produced a feasible solution with less than one second. On the other side, UMSLIC system, CELCAT requires more than one week to produce a feasible solution. Therefore, in terms of time taken to produce a feasible solution, CP performs more effective than CELCAT. By taking soft constraints into consideration, the solution will be further improved with the application of SA. For example, shifts exams with smaller size into smaller room without causing any violation of hard constraints to reduce extra room space. With the neighbourhood movement, and the acceptance of the solution based on the theory of SA, this research will produce a lower penalty cost of solution. This research runs 50 times of experiment to obtain the average result as shown in Table 4.

Table 4. Experimental Results

	Semester 2 session 2014/2015	**Semester 1 session 2015/2016**
Average Cost	6848.0	11695.6
Highest Cost	7125	11888
Lowest Cost	6357	11462
Improve (%)	21.30%	16.12%

The results in table 4 show the first dataset (semester 2 session 2014/2015) has lower average penalty cost than the second dataset (semester 1 session 2015/2016). It means the violation of soft constraints in second dataset is higher than first dataset. The different improvement rate of two datasets shows SA performed less effective in second dataset during improvement phase. This indicates the size of the dataset affects the performance of the algorithm.

However, the results show that SA is able to improve the solution from CP for both datasets. For example, the extra room capacity is calculated into penalty cost. This can be explained by when the temperature decrease, SA will only accept the solution which has lower penalty cost. However, when there is no improvement, the worse solution can be accepted under certain probability as shown in equation (10). Meanwhile, in terms of performance, the solution in first dataset is much better because there are different number of students and exams in each semester. In second dataset, more students and exams need to be scheduled and thus, increase the difficulty to schedule. This can be summarized that SA produce good solution for this examination timetabling domain. However, it does not mean SA can produce good result in other domain as stated in [21].

7 Conclusion

This research had proposed the development of Constraint Programming (CP) and Simulated Annealing (SA) to produce a better quality examination timetable. A set of formal mathematical models are developed to obtain the violations of constraints in this research. CP is developed to solve all the hard constraints without considering the violation of soft constraints. In improvement stage, SA is applied to improve the quality of solutions. Results show that SA managed to search for the better quality solution in both instances of UMSLIC.

For future research, Non-Linear Great Deluge (NLGD) can be potentially implemented in this research according to the result of [10] for course timetabling problem. NLGD provides new method to control the speed and the shape of water level decay rate which could produce perform effectively in same problem domain.

References

1. Abdullah, S.: Heuristic Approaches For University Timetabling Problems. The School of Computer Science and Information Technology (2006).
2. Abramson, D.: A Very High Speed Architecture for Simulated Annealing. 27-36 (1992)
3. Abramson, D., Krishnamoorthy, M., & Dang, H.: Simulated Annealing Cooling Schedules for the School Timetabling problem. 1-17 (1997).
4. Aycan, E., & Ayav, T.: Solving the Course Scheduling Problem Using Simulated Annealing. (n.d).
5. Cupic, M., Golub, M., & Jakobovic, D.: Exam Timetabling Using Genetic Algorithm. 357-362 (2009).
6. Groves, G., & Wijck, W.: Constraint Programming and University Timetabling. 1-12 (n.d).
7. Kuan, Y. J., Joe, H. O., & Rayner, A.: Comparison of Simulated Annealing and Great Deluge Algorithms for University Course Timetabling Problems (UCTP). American Scientific Publishers, 4, 400-407 (2015).
8. Kuan, Y. J., Joe, H. O., & Rayner, A.: A Constraint Programming Approach to Solve University Course Timetabling Problem (UCTP). American Scientific Publishers, 400-407 (2016).
9. Kuan, Y. J., Joe, H. O., & Rayner, A.: The study of Genetic Algorithm approach for university course timetabling problem. International Conference on Computational Science and Technology, 454-463 (2017).

10. Landa-Silva, D., & Obit, J. H.: Great Deluge with Non-linear Decay Rate for Solving Course Timetabling Problems. IEEE, 11-18 (2008).
11. Landa-Silva, D., & Obit, J. H.: Evolutionary Non-linear Great Deluge for University Course Timetabling. Springer-Verlag Berlin Heidelberg, 269-276 (2009).
12. Landa-Silva, D., & Obit, J. H.: Comparing Hybrid Constructive Heuristics for University Course Timetabling. 222-225 (2011).
13. Malik, A. M., Ayob, M., & Hamdan, A. R.: Iterated Two-stage Multi-neighbourhood Tabu Search Approach for Examination Timetabling Problem. IEEE, 141-148 (2009).
14. Obit, J. H., & Landa-Silva, D.: Computational Study of Non-linear Great Deluge for University Course Timetabling. Springer-Verlag Berlin Heidelberg, 309-328 (2010).
15. Obit, J. H., Landa-Silva, D., & Ouelhadj, D.: Non-Linear Great Deluge with Learning Mechanism for the Course Timetabling Problem. Springer-Verlag Berlin Heidelberg, 1-10 (2009).
16. Ozcan, E., Misir, M., Ochoa, G., & Burke, E. K.: A Reinforcement Learning - Great-Deluge Hyper-Heuristic for Examination Timetabling . 1-25 (n.d.).
17. Rankhambe, M., & Mrs.S.Kavita.: Optimization of Examination Timetable Using Harmony Search Hyper-Heuristics (HSHH). Internation Journal of Computer Science and Information Technologies, 5, 6719-6723 (2014).
18. Socha, K., Knowles, J., & Samples, M. (2002). A MAX-MIN Ant System for the University Course Timetabling Problem. Springer-Verlag London, 1-13 (2002).
19. Tuan-Anh Duong, Kim-Hoa Lam.: Combining Constraint Programming and Simulated Annealing on University Exam Timetabling. 205-210 (n.d.).
20. Valouxis, C., & Housos, E.: Constraint programming approach for school timetabling. Elsevier Science Ltd., 1555-1572. (2003).
21. Wolpert, D. M.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 67-82 (1997).
22. Wong, T., Cote, P., & Gely, P.: Final Exam Timetabling: A Practical Approach. IEEE, 726-731 (2002).

Time Task Scheduling for Simple and Proximate Time Model in Cyber-Physical Systems

Yuan FANG[1][2], Sian En OOI[1], Yuto LIM[1], and Yasuo TAN[1]

[1] Japan Advanced Institute of Science and Technology (JAIST),
1-1 Asahidai, Nomi, Ishikawa, 923-1292 JAPAN,
{yfang, sianen.ooi, ylim, ytan}@jaist.ac.jp,
WWW home page: http://www.jaist.ac.jp
[2] Dalian Polytechnic University (DPU),
No.1 Qinggongyuan, Dalian, Liaoning, CHINA

Abstract. Modeling and analysis play essential parts in a Cyber-Physical Systems (CPS) development, especially for the system of systems (SoS) in CPS applications. Many of today's proposed CPS models rely on multiple platforms. However, there are massive reusable components or modules in the different platform. And also, the model had to be modified to meet the new system requirements. Nevertheless, existing time model technologies deal with them, but it leads to a massive time consuming and high resource cost. There are two objectives in this paper. One is to propose a new simple and proximate time model (SPTimo) framework to the practical time model of hybrid system modeling and analysis. Another is to present a time task scheduling algorithm, mix time cost and deadline first (MTCDF) based on computation model in the SPTimo framework. Simulation results demonstrate that the MTCDF algorithm achieves the priority the scheduling of tasks with a time deadline, and match with optimal scheduling in time requirement and time cost.

Keywords: cyber-physical systems, time model, scheduling, deadline first, optimal scheduling index

1 Introduction

Cyber-Physical Systems (CPS) in [1] and [2] is usually defined as tight integration of computation, communication, and control with deep interaction between physical and cyber elements in which embedded devices, such as different sensors and actuators, are wireless or wired networked to sense, monitor and control the physical world. With CPS becoming more popular, many types of research have devoted to their development. The modeling and analysis play an essential part of the safety and mission critical system of systems (SoS) development in CPS. Researches in [3] have contributed to the modeling of CPS. Some of the modelings of the discrete and continuous behavior of heterogeneous systems has been developed recently, such as Ptolemy II [4] and Programming Temporally Integrated Distributed Embedded Systems (PTIDES)[5], those models for

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_19

a deterministic modeling paradigm suitable for CPS application at any scale. All most of the current modeling structure is platform to platform, it follows discrete-event (DE) semantics. Furthermore, time is the first parameter because it is the interaction requires physical world and cyber world in CPS.

The model-driven development method first needs to solve the structure of the model, then to solve the task scheduling. Existing CPS modeling method is the procedure and initial process of the modeling are still highly complicated. Furthermore, time is very consuming because it requires a full and in-depth understanding of the details of the physical environments. Its the weakness which it is difficult, even impossible, to adapt some existing scheduling algorithms different systems or to different scheduling targets. When we design a scheduling algorithm for a newly developed system, it is difficult to benefit from the existing scheduling algorithms, which usually results in a high cost. For many existing CPS scheduling algorithms, to enhance them to support some practical requirement is difficult, even impossible. Due to system complexity, simple and proximate modeling purpose method for CPS application.

The objectives of this paper are to propose a simple and proximate time model (SPTimo) framework for CPS applications to meet the requirements of modeling and analysis with minimum cost and time firstly. This model based on a proximity method, which is used to model each component of sensing, fabric network, and actuating in the reality multiple platforms model as computation model, control model, and communication model, respectively. Second, Mixed Time Cost and Deadline-First (MTCDF) algorithm was proposed to addressed the problem of multi-programming scheduling in the SPTimo computation model. At last, the successful ratio and optimal scheduling index were shown in the simulation results.

2 Related Works

2.1 Modeling of CPS

CPS modeling requires portrayal of how interactions between the process and physical processes are calculated and how they behave when they are merged [5]. The model-based analysis provides a better understanding of CPS behavior, and model-driven design can improve design automation and reduce errors in refinement. Edward Lee proposed CPS is integrated computing power and physical process of the system[6], which uses embedded computers and networks to monitor the physical processing process, and with the feedback loop, physical process and computation process affect each other. For a system modeling, simulation and verification, computational science and control science have different ways. However, in the control science, the research on the physical world is often based on time, abstracting the system as a continuous time model, and time is the most critical factor in the model this will result in collisions and random failures in the interaction between the computational unit and the physical entity model. There are various interactive entities involved in CPS system. It can be a natural environment, building, machine, a physical device, human beings, etc. The situation can be real-time sensing part (such as the physical entity) can also be controlled.

The current reality model of CPS in Fig. 1, most of them are multiple platforms structure. There are sensors, actuators, computation units in the platform. If there is a *Service* need all the platforms 1, 2 and 3. The *Service* will be in computation time $\sum_{i=1}^{4} com_i$, and the all the delay is $\sum_{i=1}^{5} d_i$. There is two main problem with this model: *i)* the platform usually uses a single algorithm to schedule. For services with different time requirements, time efficiency is not optimal.*ii)* some computation time is repeated. From the perspective of the entire system, different computing capabilities have different effects on the time of service. Accumulating these effects together will cause the time and task of the next service to miss deadlines and cannot be implemented, which reduces the efficiency of the system.

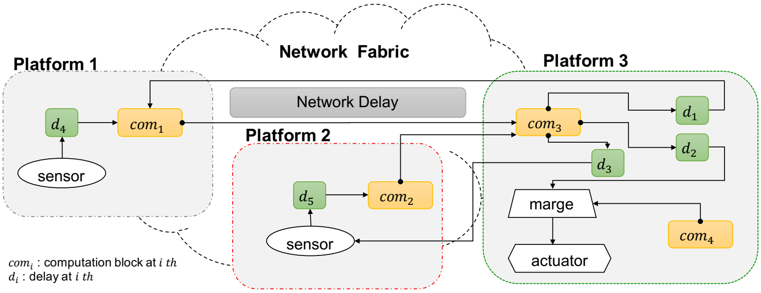


Fig. 1: General time model for CPS

2.2 Time Model

Discrete-Event (DE) is used to model time, discrete interactions between concurrent actors. [6] There *event* is in each communication, conceptually understood to be an instant message sent from one actor to another. The ***time model*** has encompassed the overall approach to handling time sequencing.

3 Simple and Proximate Time Model Frameworks

3.1 Structure of the SPTimo Framework

The SPTimo framework is illustrated in Fig. 2. There are three sub-models included in the frame. The first sub-model is ***Computation Model***. All the components with computational requirements from the different platform are approximated by this model. A service request requires connecting different platforms in a specific order. Each platform has different devices that complete the tasks. The operation generated by each device is called a *Time Task*. In a CPS service, time tasks are required to be completed before their deadlines. To satisfy this requirement, their required computation resources should be allocated to tasks at the right time. The allocated result is called *schedule*, while the allocating process is called *scheduling* which is conducted by a scheduler equipped in the system. The second sub-model is ***Control Model***. All the components of the approximate control operation are integrated into this model. The control model is mainly proposed in the control model, especially the feedback control algorithm, which includes online and offline forms. ***Communication Model*** follows the existing communication protocols. It mainly includes the network synchronization unit and the time offset of the SPTimo framework.

In fig. 1, there are four computation units, and they will feedback the result to the other parts. If the case in the SPTimo framework, there is one computation time and get the optimal schedule for one service.

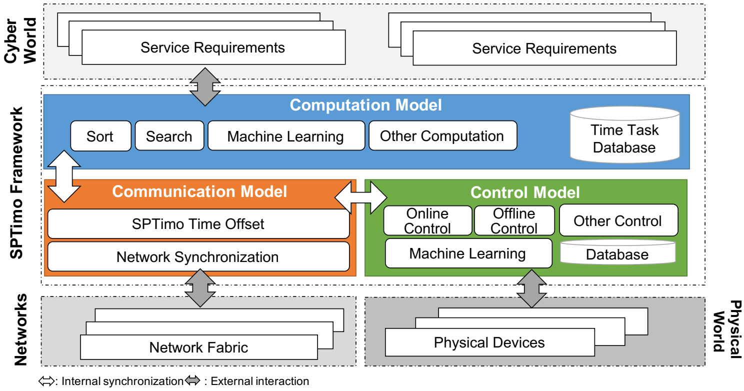


Fig. 2: SPTimo framework

3.2 Time Task Model

Definition 1 (Time Task Dependency Graph) A time task dependency graph is a directed non-cyclic graph. $G = (P, E)$, where P is a platform set, $E \subseteq P \times P$ is a dependency realation (edge) set, with $(p_i, p_j) \subset E$,$i \neq j$,where $p_i, p_j \subset P$. An edge (p_i, p_j) in the task dependency graph means platform p_i can start to execute only after platform p_i has been completed. $p_i \prec p_j$ is used to illustrate this dependency relation, and the this relation is transitive.

Definition 2 (Time Task) A time task is a multidimensional set $TT = (T, S, V)$. It includes a subset of time T, a subset of states S, and a subset of values V. The j time task of i platform is $T_{i,j} = \{A_{i,j}, E_{i,j}, D_{i,j}\}$, $A_{i,j}$ is the arrival time in a task scheduling, $E_{i,j}$ is the execute time, and $D_{i,j}$ is the deadline. Each task $S_{i,j} = \{s_{i,j}\}$, $s_{i,j}$ is the state of the elements. $V_{i,j}$ is value of the time task $T_{i,j}$ with state $S_{i,j}$. i is means the i time task, and j is means the j platform, $(i, j) \subseteq N$.

Definition 3 (Service) A service is defined by the operation of the platforms with the order of execution in the CPS application. $Service = (P_i \prec P_j)$. We assume that the execution time task $T_{i,j}$ inside the platform P_i are disordered. In order to meet the highest proportion of successful service execution, the time task sequence within the platform can be adjusted.

Eqn. (1) is used to calculate the waiting time of the time task. Equ. (2) for determining whether the time tasks can be executed.

$$W\left(_{i,j}\right) = A_{i,j} + \sum E\left(_{i',j'}\right) \tag{1}$$

where i', j' is means the total task before i, j.

$$D\left(_{i,j}\right) \geq W\left(_{i,j}\right) \tag{2}$$

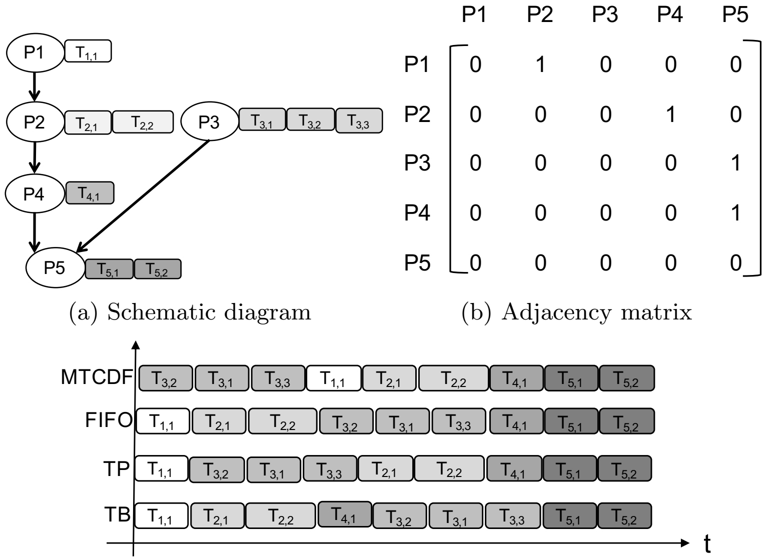
$\forall T\left(_{i,j}\right)$ satisfied the $Equation\,(2)$, the service entire time task is executable.

4 Time Task Scheduling

4.1 Scheduling Procedure

There are many scheduling algorithms used in various real-time systems. In this subsection, through an example described in Fig. 3, the performance of three widely used scheduling algorithms, Tree Based (TB), Task Proirity (TP), and First In First Out (FIFO) are studied. In the case, the lengths of the indivisible fragments in the tasks are shown the execute time length. The parameters of the platform time task are shown in Table 1. Schedules the time task with the direct way according to the following adjacency matrix. Each row corresponds to the starting point, and each column corresponds to the ending point. For example, $P_1 \prec P_2$, the intersection of the first row and the second column has a value of 1 in the adjacency matrix.

Dependency the adjacency matrix, fig. 3 (b), Table 2 shows that the four scheduling results.


P1
$T_{1,1}$
P2
$T_{2,1}$
$T_{2,2}$
P3
$T_{3,1}$
$T_{3,2}$
$T_{3,3}$
P4
$T_{4,1}$
P5
$T_{5,1}$
$T_{5,2}$

(a) Schematic diagram

$$\begin{array}{c|ccccc} & P1 & P2 & P3 & P4 & P5 \\ \hline P1 & 0 & 1 & 0 & 0 & 0 \\ P2 & 0 & 0 & 0 & 1 & 0 \\ P3 & 0 & 0 & 0 & 0 & 1 \\ P4 & 0 & 0 & 0 & 0 & 1 \\ P5 & 0 & 0 & 0 & 0 & 0 \end{array}$$

(b) Adjacency matrix

MTCDF $T_{3,2}$ $T_{3,1}$ $T_{3,3}$ $T_{1,1}$ $T_{2,1}$ $T_{2,2}$ $T_{4,1}$ $T_{5,1}$ $T_{5,2}$
FIFO $T_{1,1}$ $T_{2,1}$ $T_{2,2}$ $T_{3,2}$ $T_{3,1}$ $T_{3,3}$ $T_{4,1}$ $T_{5,1}$ $T_{5,2}$
TP $T_{1,1}$ $T_{3,2}$ $T_{3,1}$ $T_{3,3}$ $T_{2,1}$ $T_{2,2}$ $T_{4,1}$ $T_{5,1}$ $T_{5,2}$
TB $T_{1,1}$ $T_{2,1}$ $T_{2,2}$ $T_{4,1}$ $T_{3,2}$ $T_{3,1}$ $T_{3,3}$ $T_{5,1}$ $T_{5,2}$
t

(c) Scheduling algorithms results

Fig. 3: Example for scheduling

Table 1: Time task parameters

Platform	$T_{i,j}$	$A_{i,j}$	$E_{i,j}$	$D_{i,j}$
P1	$T_{1,1}$	0	1	5
P2	$T_{2,1}$	1	1	10
	$T_{2,2}$	1	4	10
P3	$T_{3,1}$	0	1	5
	$T_{3,2}$	0	1	1
	$T_{3,3}$	0	1	5
P4	$T_{4,1}$	2	1	10
P5	$T_{5,1}$	3	1	15
	$T_{5,2}$	3	1	15

Table 2: The scheduling result of example

Algorithm	*Service* Scheduling	$\sum_{i}^{\forall j} W\left(_{i,j}\right)$	Successful Ratio	Unable to Meet Deadline
TB	$\{P_1 \prec P_2 \prec P_4 \prec P_3 \prec P_5\}$	44	0.6667	$T_{3,1}$, $T_{3,2}$, $T_{3,3}$
TP	$\{P_1 \prec P_3 \prec P_2 \prec P_4 \prec P_5\}$	35	0.8889	$T_{3,2}$
FIFO	$\{P_1 \prec P_2 \prec P_3 \prec P_4 \prec P_5\}$	44	0.6667	$T\ _{3,1}$, $T_{3,2}$, $T_{3,3}$
MTCDF	$\{P_3 \prec P_1 \prec P_2 \prec P_4 \prec P_5\}$	44	1.0000	None

Algorithm 1 Mixed Time Cost and Deadline First (MTCDF)

```
A $Service$ is processing schedule defined by an adjacency matrix $|AJ\,(P_i \prec P_j)|$,$P_i$,$P_j$ is platform, for all $i, j, k \subseteq N$
$S_{MTCDF} = AJ.\texttt{MTCDFSort}(T_{i,j})$
for all $1 \le i \le n$ do
    for all $1 \le j \le n$ do
        $sum_i = T_{i,j}^{deadline}$
        sort($T_{i,j}$) with deadline first
    end for
end for
for all $1 \le i \le n$ do
    if $sum_i <= sum_{i+1}.and.P_i.AJ == P_{i+1}.AJ$ then
        $S_{k++} = P_i$
    else
        $S_{k++} = P_{i+1}$
    end if
end for
Scheduling list is $S_{MTCDF} = list(S_k, 1 \le k \le n)$
```

4.2 Mixed Time Cost and Deadline-First (MTCDF) Algorithm

Dynamic scheduling of tasks in an overloaded real-time system was proposed by [7]. Cheng et al. propose SMT (*satisfiability modulo theories*)-based scheduling method [8]. Lim et al. propose time delay model for smart home [9]. Those pursues to focus on maximizing the total number of tasks that can be completed before their deadlines and time delay model with experiment. In the paper, we propose one navel scheduling algorithm, mixed time cost and deadline first (MTCDF) base the time task database of SPTimo Framework. This method is used to priorities the scheduling of tasks with a time deadline, and records the computation time.

Some assumptions were applied to the proposed scheduling algorithms, e.g., *1*) The requests for all time tasks for which strict deadlines exist are random; *2*) Time tasks are independent for each platform; *3*) Run-time for each time task is constant and does not vary with time.

According to the time task example, there are two main parts that affect the efficiency of the algorithm.*i*) The order of execution of the platforms in the same priority situation. *ii*) The time cost to compute the time task schedule in the deadline-first platform. Such as the scheduling $T_{3,2} \prec T_{3,1} \prec T_{3,3}$.
The relationship between deadline-first and time cost is defined by the following equation:

$$\varGamma_k = \mu \cdot \sum_{i=1}^{n} P_i^{deadline} + \nu \cdot \sum_{j=1}^{m} T_{i,j}^{timecost} \tag{3}$$

where $\varGamma$ is means the optimal scheduling index, k is the algorithm, μ is the coefficient of deadline-first part, ν is the coefficient of time cost part. $i, j, n, m \subseteq N$. The process of schedule synthesis is summarized in Alg. 1.

5 Simulation and Results

5.1 Simulation

In order to evaluate the successful ratio among the TP, TB, FIFO, and MTCDF scheduling algorithms in different λ and service including platforms' number. The input *Service* is generated according to the adjacency matrix AJ. There is one time task database, with 10 times number of platform random time tasks. The input functions random distribute the time tasks to platform from time task database. Each time task has three parameters, in the Def. (2). For each arrival time $A_{i,j}$ is generated according to Poisson distribution with arriving rate λ to every platform. In this simulation, the average inter arrival time of task is fixed at 100 ms per frame, where as the average inter arrival time of task varies from 10 to 1000 ms per frame. All the simulation results is collected using a 64bit Intel Core i7 CPU 2.4 GHz with 16GB of memory.

5.2 Results

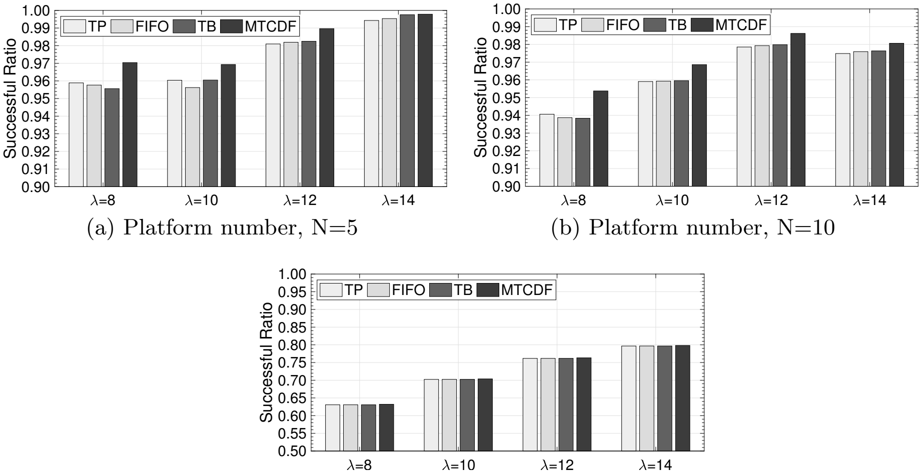
A service is performed by different time tasks of multiple platforms according to the scheduling result. In this simulation, the number of time tasks in the time task database is 10 times the number of platforms, and randomly matched to the platform as the service is generated. The arrival time and deadline of the time task are with the Poisson distribution. With $\lambda = \{8, 10, 12, 14\}$, the time task number is the the successful ratio and the optimal index $\mathit{\Gamma}_k$ was observed. The results presented in Fig. 4 and Fig. 5 are averaged over 100 simulation runs with different random numbers.

Successful Ratio We have measured the effect of successful ratio on different platform number with different λ. As seen from Fig. 4 (a), when the λ is changed, the successful ratio are increased. Especially, more MTCDF can improve more succesful ratio in the same λ comparing with TP, TB, and FIFO. As Fig. 4 (b), we show the successful ratio with 10 platform. In Fig. 4 (c), we show the 100 platform. As the result, the successful ratio is over 0.648. And the four scheduling algorithms results are very close.

Optimal Scheduling Index The optimal scheduling index refers to the more appropriate scheduling was found. We use the min-max normalization method to process the results. In Fig. 5 shows the optimal scheduling index $\mathit{\Gamma}$ with different λ different platform number N. The results show that the optimal scheduling index can give to optimization scheduling results, with meeting the time requirements first and time cost.

6 Conclusion and Future Work

In this paper, we have addressed the platform to platform scheduling issues of the CPS time model. We define a new time model the SPTimo framework in

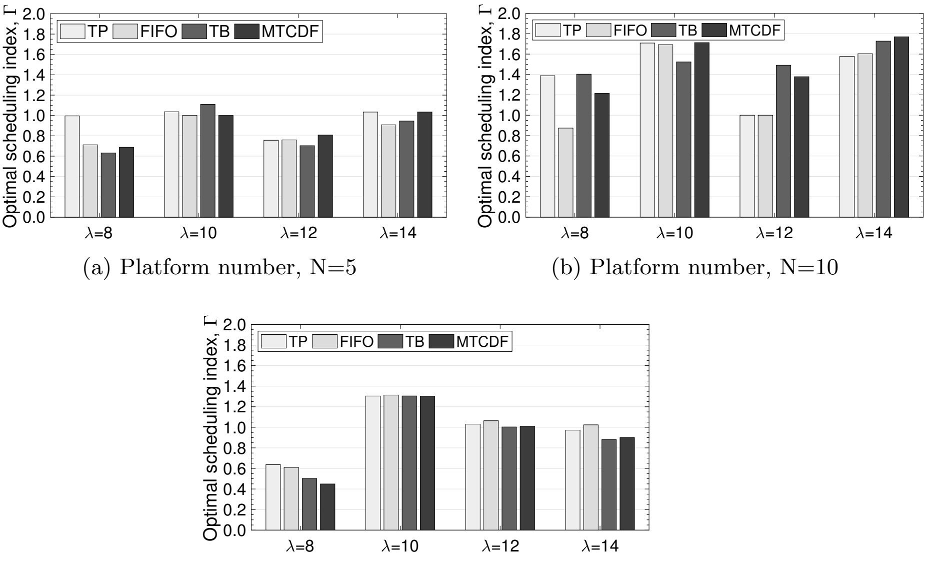


(a) Platform number, N=5

(b) Platform number, N=10

(c) Platform number, N=100

Fig. 4: Successful ratio



(a) Platform number, N=5

(b) Platform number, N=10

(c) Platform number, N=100

Fig. 5: Optimal scheduling index, Γ

CPS. The contributions of the SPTimo framework are it reduce the repetition rate of the components and the elements in the same model can be approximated, to enhance the efficiency of computation and control. According to the SPTimo framework design principle, we proposed the MTCDF algorithm for computation model scheduling algorithm. The simulation results show that the MTCDF time task scheduling method can improve service success rate. The SPTimo framework can generality of the matching optimal scheduling index for time requirement service.

As future work, we will extended the machine learning algorithm of computation model in SPTimo framework. Then we will integrate the control model to SPTimo framework.

References

1. E. A. Lee, "CPS foundations," *ACM/IEEE Des. Automation Conf. (DAC)*, pp.737742, 2010.
2. B. Sean, S. Shankar, A. John, S. Roundtable, "Reliability of embedded and cyber-physical systems," *IEEE Security and Privacy*, vol.8,no.5 : pp27-32, 2010
3. J. C. Eidson, E. A. Lee, S. Matic, S. A. Seshia and J. Zou, "Distributed Real-Time Software for CyberPhysical Systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 45-59, Jan. 2012.
4. P. Derler, E.A. Lee, and A.S. Vincentelli, "Modeling cyber-physical systems," *Proceedings of the IEEE*, vol.100, no.1, pp.1328, 2012.
5. P. Derler,T.H. Feng,E.A. Lee, et al. "PTIDES: A programming model for distributed real-time embedded systems," *California Univ Berkeley Dept of Electrical Engineering and Computer Science*, 2008.
6. E. A. Lee, "The Past, Present and Future of Cyber-Physical Systems: A Focus on Models," *Sensors (Basel, Switzerland)* , vol.15, no.3, pp. 4837-4869. PMC. Web, 2017.
7. A. Marchand, M. Chetto, "Dynamic scheduling of periodic skippable tasks in an overloaded real-time system," *IEEE/ACS International Conference on Computer Systems and Applications*, Doha, Qatar, pp. 456464, Apr. 2008.
8. Z. Cheng, H. Zhang,Y. Tan, et al. "SMT-Based Scheduling for Overloaded Real-Time Systems," *IEICE Transactions on Information and Systems*, 2017, 100(5): 1055-1066.
9. Y. Lim, Y. Tan. "Time Delay Modeling for Energy Efficient Thermal Comfort Control System in Smart Home Environment," *International Conference on Computational Science and Technology.* Springer, Singapore, 2017: 42-52.
10. P. Palensky, E. Widl, A. Elsheikh. "Simulating cyber-physical energy systems: Challenges, tools and methods," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2014, 44(3): 318-326.
11. A. Murugesan, S. Rayadurgam, M.W. Whalen, et al. "Design Considerations for Modeling Modes in Cyber-Physical Systems," *IEEE Design & Test*, 2015, 32(5): 66-73.
12. M.D. Ilic, L. Xie, U.A. Khan, et al. "Modeling of future cyberphysical energy systems for distributed sensing and control," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2010, 40(4): 825-838.

DDoS Attack Monitoring using Smart Controller Placement in Software Defined Networking Architecture

Muhammad Reazul Haque[1], Saw C. Tan[1], Zulfadzli Yusoff[2] , Ching K. Lee[2], Rizaludin Kaspin[3]

[1] Faculty of Computing & Informatics, Multimedia University, Jalan Multimedia, Cyberjaya 63100, Malaysia
[2] Faculty of Engineering, Multimedia University. Jalan Multimedia, Cyberjaya 63100, Malaysia
[3] Telekom Malaysia Research and Development, TM Innovation Centre,Lingkaran, Cyberjaya 63000,Malaysia
reazulh@gmail.com, sctan1@mmu.edu.my, zulfadzli.yusoff@mmu.edu.my, cklee@mmu.edu.my, rizaludin@tmrnd.com.my

Abstract. Software defined networking is upcoming agile networking system for computer and telecommunication. The apartness of data plane and control plane are key points of SDN architecture which enables flexibility and programmability of network. But distributed denial of service attack is the main threat for software defined networking architecture as it can send huge traffic directly to the controller. In this paper, we proposed a hypothetical concept of smart controller placement for SDN architecture which can monitor controller's health status and provide continuous services during the DDoS attack.

Keywords: SDN, Smart Controller placement, DDoS Attack, Software Defined Networking, Controller of Controller, DDoS attack monitoring.

1 Introduction

Software defined networking (SDN) has enticed many researchers and scientists around the world because it is financially savvy, centralized, programmable, magnificent, superb and upcoming networking systems. Traditional computer and telecommunication networks are complex and difficult to manage. One of the alarming issue related to SDN security is its vulnerability to attacker which can send huge amount of distributed denial of services (DDoS) attack traffic to the controller of SDN [1]. In SDN if controller becomes the prey of DDoS attack, all the switches linked to that controller will be malfunctioned and fail to continue the operation. As a result, SDN service for the legitimate user will be obstructed. So the DDoS onfall is the rampant threat for subsequent SDN [2]. In this paper, we propose a smart controller placement model which can ensure the network operation under DDoS attack. Smart controllers (SC) are powerful controller of controllers.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_20

In OpenFlow [3], the switches will typically ask controller to get new flow rules for any data flow which they do not know how to forward [2]. SDN Switch only check for match in forwarding table, when there is no match it will directly send packets to controller to process [4]. Controller is the brain of SDN process and decides where to forward that packet. DDoS attacker can send massive amount of spoofed data packet as this type of fake packet are not harmonized by the flow table of switch. So controller will be engage to process these huge fake traffic of new incoming packets and creates similar flow entries harder better faster stronger, also over and above newly installed spoofed flow entries will grab the full flow table in the openflow switch immensely soon [5], which may lead to failure of the SDN controller [2].

Security analysis reported that the framework of SDN tolerates numerous security dangers, together with [5] : 1) disallowed access like unauthenticated controller and application access 2) leakage of data like flow rule hacking, forwarding policy hacking, data packet processing technology timing analysis, 3) data exchange like flow method exchange to modify data packets, 4) unpleasant or bitchy applications like deceitful rule addition controller seizing 5) configuration problem like lack authentication techniques and 6) DDoS like switch-controller communication surge and switch drift table flooding. So the smart controller's characteristic needed to face the above technical security issue.

SDN controllers are at the focal point of the SDN engineering architecture, mediating between clients and resources to deliver services [6]. In [7] reported that there is no single SDN controller that incorporates every one of the distinguished features for safe, powerful, and resilient services. A model to give nonstop SDN benefit under DDoS assault by the commands sent by the controller and smart savvy controller placement is essential.

Open Networking Foundation (ONF) [8] summarized that Security, Flexibility, Distributed controller considerations, Controller deployment, Interworking with non-SDN environments, Application-controller plane interface capabilities, Network initialization, Protection and restoration, Logging, particularly security logs. Log stockpiling, transfer and maintenance, Configuration and administration industriousness, reinforcement, reclamation, reviewing, Traffic investigation, network and equipment planning, installation, establishment, stock features are required for SDN controller [9].

Smart controller required some more features like controller's health monitoring, low traffic attack [2] analysis, packet process sharing etc. In section II and III we discussed about SDN and DDoS attack. Section IV presents the DDoS attack propagation in SDN. We furnished the Smart controller placement model in section V and finally we conclude this paper in section VI.

2 Software Defined Networking (SDN)

Software Defined Networking (SDN) is a rising design that is dynamic, sensible, practical, and versatile, making it perfect for the high-data transfer capacity, dynamic nature of the present applications. This design decouples the system control and forwarding capacities [8]. The fundamental principle of SDN is to disjunction of the control plane and the data plane [10] by developing special operating system software that can run on hetero hardware [11]. This disjunction brings superb features for future network but at the same time it will be a threat for SDN security [12].

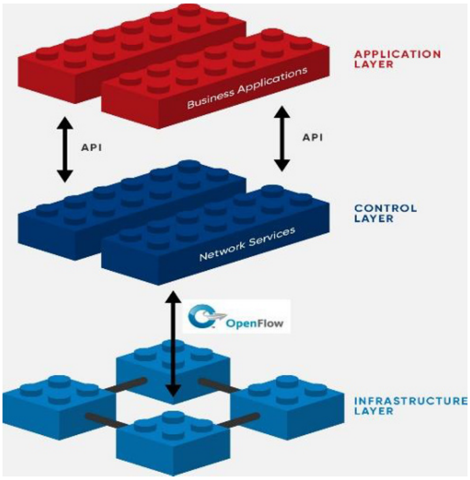


Fig. 1. SDN Architecture with 3 layers [8].

There are three layers in SDN architecture such as application layer, control layer and infrastructure layer or data layer. SDN has been proposed as a postulant of the future network and telecommunication architecture, SDN already adopted in internal data centres by giant companies like Google [13]. SDN designed for expediting and progressing network management system with high flexibility and reliability by rending control plane and data plane. More Innovation opportunities tern leads by network programmability of SDN. Numerous researchers' attention from both industrial and academic has attracted for SDN issues. [14].

3 DDoS Attack

An endeavor to make an online administration or system or SDN unattainable by overloading with huge spoofed data packet traffic from abundant sources is called distributed denial of services (DDoS) attack. Smartphone, PC, Alarm system, Server, smart camera or any web associated gadgets like internet of things (IoT) can be the genesis of DDoS assault. It's simple to perform by sending thousands of botnets or zombai but difficult to trace the attacker. Now a day anyone can buy or rent 50,000 botnets for $1000 only or $9 per hour [15, 17]. Botnets are attacking armies that formed by bots or zombai. Srizbi, Kraken/ Bobax, and Rustock are large-scale botnets picked up reputation for their sizes and pernicious activities [16]. Attacker are presently leasing access to a monstrous Mirai botnet, which they guarantee has in excess of 400,000 tainted bots, prepared to do DDoS attack at anybody's command [17].

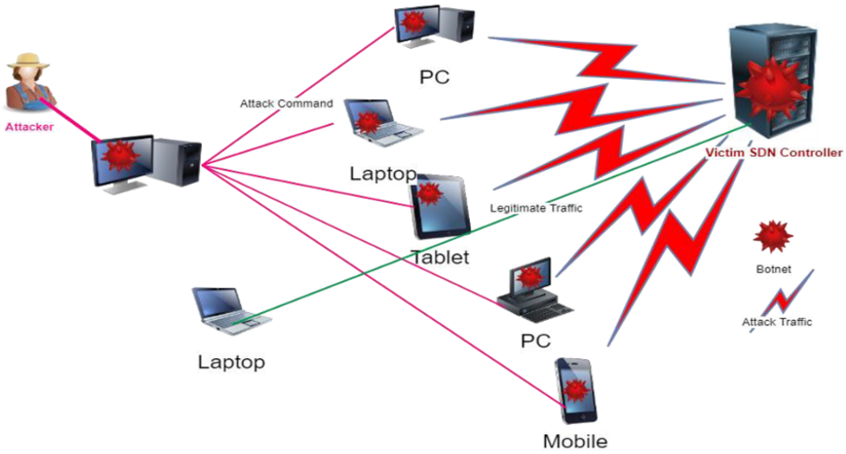


Fig. 2. DDoS attack using botnet.

We illustrate DDoS attack utilizing botnet in Fig. 2 An aggressor send botnets to several internet connected devices like PC, tab and smart phone. After that he commands all botnet to send huge traffic at a time to SDN controller through switch. We see the switch and controller becomes malfunctioned after attack. As a result all the devices connected with that victim controller will be malfunctioned too. This is how a clever attacker can interrupted SDN services for legitimate user.

4 DDoS Attack Propagation through SDN Architecture

In Fig 3 illustrated that attacker can send DDoS attack packet (DAP) to data plane or infrastructure layer via routers, switches, wireless access points etc. These huge DAP can propagate to the control layer through southbound API like Openflow, ForCES Netconf etc. DAP contains vast amount of fake traffic to flood controller, flow table of switch and any application. In control layer the DAP can spread to controller to controller through westward and eastward API like ALTO, Hyperflow etc. [17].

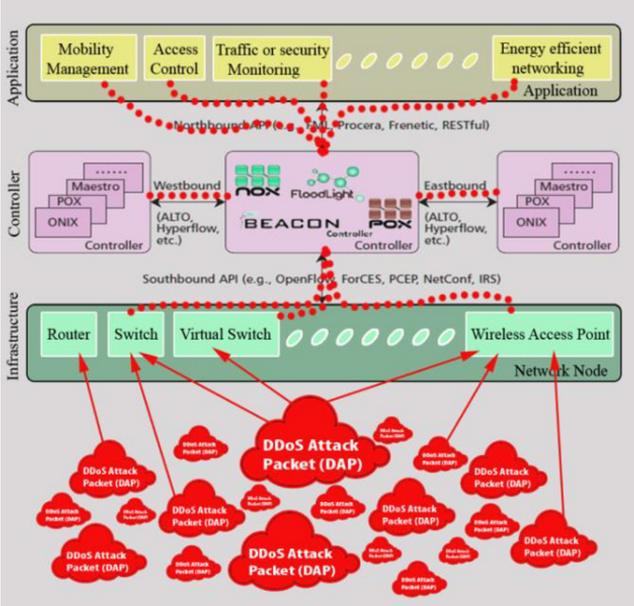


Fig. 3. DDoS attack on SDN Architecture. [18]

Finally these spoofed traffics can propagate to the application layer, for example, portability administration, get to control and activity or security checking via northbound API like RESTful to disable SDN services.

5 SDN Smart Controller Placement

Controller is the cerebrum of SDN. Some issues of controller placement like how many controller and where to deploy controller already well addressed but the issue of controller placement in security threat like DDoS attack is yet to address. If controller becomes the victim of DDoS attack, all the switches connected to that controller will be malfunctioned and unable to serve the legitimate user. DDoS attack is easy to perform but very hard to detect. In this sense we must install alternative controller to serve the legitimate user. In the next section we will describe about smart controller placement which will lead to continuous SDN services.

Smart controller is a powerful controller of controller. The processing power, the number of port, the available controller and the price will be higher than controller. If any controller becomes malfunctioned after receiving DAP, the smart controller will be active and start work. It will take all responsibility of controller to continue the services of SDN. This concept will be a great threat for DDoS attacker.

Here, we introduce DDoS attack aware controller placement which consists of extra smart controllers in addition to the existing controllers to ensure the serve to legitimate users without severance.

In SDN if switch do not get any match in forwarding flow table it send request to controller via openflow to create rule to process that packet. Similarly a smart controller can place to monitor all controller activities. Especially for DDoS attack traffic monitoring. If any controller become congested to process fake traffic it should sent request to smart controller to create rule or share with other controller.

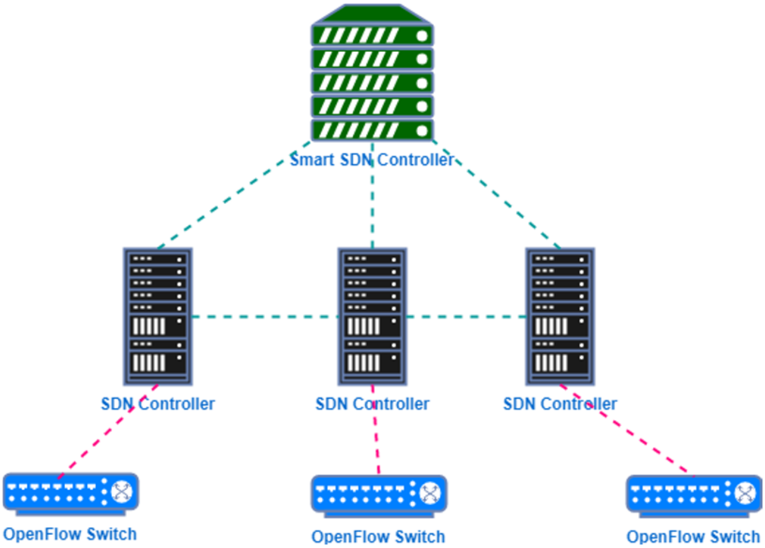


Fig. 4. Smart Controller placement to monitor controllers

Normally the smart controller will not process any data from the switch. Mainly, a smart controller will take care of all controllers' health and security issue. It will apply rule to controllers only.

Smart controller should be placed in the upper bound part of control plane of SDN architecture. We have discussed about functions and features of the smart controller. It is essential to place smart controller where high level of network security is required.

Table. 1. Controller and Smart Controller's functions.

Function	Controller	Smart Controller
Packed Forward to Switches	Yes	No
Link to Switches	Yes	No
Share data packet	No	Yes
Monitor Controller	No	Yes
SDN Failure Risk	High	Low

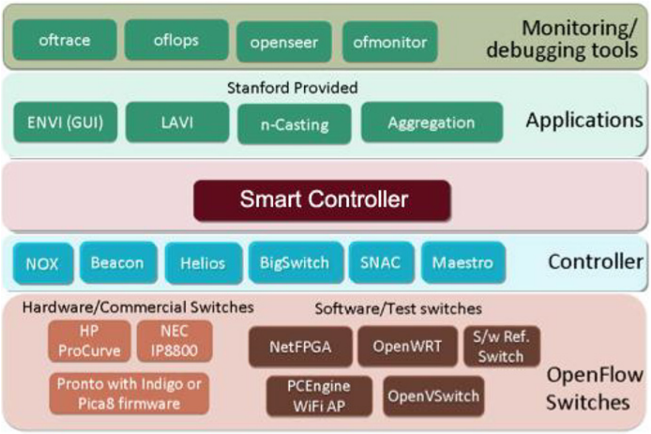


Fig. 5. SDN architecture and Smart Controller placement concept

We illustrated a smart controller concept on SDN architecture [19] in Fig 5. Here is the basic flow chart of smart controller in SDN.

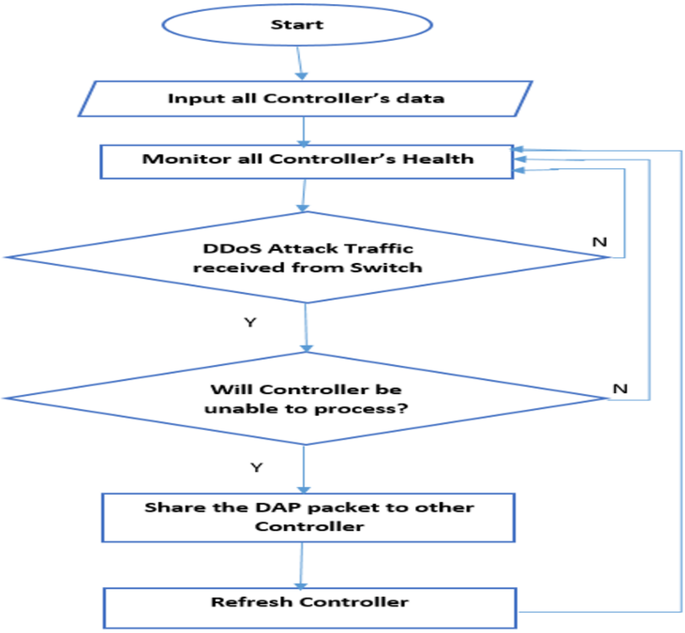

Start
Input all Controller's data
Monitor all Controller's Health
DDoS Attack Traffic received from Switch
N
Y
Will Controller be unable to process?
N
Y
Share the DAP packet to other Controller
Refresh Controller

Fig. 6. Flow chart of Smart Controller

Smart controller will be able to refresh all controllers and send command to share the DAP packet to other controller to provide continues SDN service.

6 Conclusion and Future Work

The proposed hypothetical smart controller placement model is capable to provide continuous SDN service. It will monitor the condition of health of all controllers. Smart controller can forward packet request to other controller to process. During the DDoS attack it can distribute the packet request to different controller. Smart controller is the controller of controller. This concept is agile enough for highly secured network. In future we will design the smart controller functions using artificial intelligence (AI), artificial neural network (ANN) and machine learning (ML) technology for SDN architecture. It depends on vast research.

Acknowledgment

This research work is completely upheld by Telekom Malaysia (TM) R&D (Project SAP ID: MMUE/160017.02) and Multimedia University (MMU), Cyberjaya, Malaysia. We are extremely grateful to the group of TM R&D for giving the help to our exploration thinks about.

References

1. Scott-Hayward, S., Natarajan, S., & Sezer, S. A Survey of Security in Software Defined Networks. IEEE Communications Surveys and Tutorials, 18(1), 623-654, 2016, DOI: 10.1109/COMST.2015.2453114
2. P. Dong, Xiaojiang, H. Zhang, T. Xu, A Detection Method fo a Novel DDoS against SDN Controller by Vast new Low-Traffic flows, IEEE ICC Communication and Information Systems Security Symposium, 2016
3. N. McKeown et al., OpenFlow: Enabling innovation in campus networks & quot;. ACM SIGCOMM Computer Communication Review archive Volume 38 Issue 2, April2008 Pages 69-74. Doi: 10.1145/1355734.1355746, Accessed March 2018: https://dl.acm.org/citation.cfm?id=1355734.1355746
4. S. M. Mousavi and Marc St-Hilaire, Early Detection of DDoS Attacks against SDN Controllers, IEE International Conference on Computing, Networking and Communications, and Information Security Symposium, 2015
5. Q. Yan, F. Richard Yu, Senior Member, IEEE, Qingxiang Gong, and Jianqiang Li, Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 1, FIRST QUARTER 2016.
6. S. Schaller, D.Hood, Software defined networking architecture standardization, Computer Standards & Interfaces, Volume 54, Part 4, 2017, Pages 197-202, ISSN 0920-5489, https://doi.org/10.1016/j.csi.2017.01.005.
7. S. Scott-Hayward, Design and deployment of secure, robust, and resilient SDN Controllers, IEEE, 2015
8. Open Networking Foundation (ONF), Software-Defined Networking (SDN) Definition, 2018, Accessed on 9 June 2018: https://www.opennetworking.org/sdn-definition/
9. M. Betts, N. Davis, R. Dolin, P. Doolan, SDN architecture, Open Networking Foundation, Issue 1, June, 2014.

10. P., Xia., L. Zhi-yang, G. Song, Q. Heng, Q. Wen-yu, Y. Hai-sheng. AKself- adaptive SDN controller placement for wide area networks, Frontiers of Information Technology & Electronic Engineering, ISSN 2095-9184 (print); ISSN 2095-9230 (online). 2016.
11. M.Seliuchenko, Orest Lavriv, Oleksiy Panchenko, Volodymyr Pashkevych, Enhanced Multi-commodity Flow Model for QoS-aware Routing in SDN, 2016 International Conference "Radio Electronics & InfoCommunications" (UkrMiCo) September 11-16, 2016, Kiev, Ukraine.
12. Q. Yan and F. Richard Yu , SECURITY AND PRIVACY IN EMERGING NETWORKS, Distributed Denial of Service Attacksin Software-Defined Networking with Cloud Computing, IEEE Communications Magazine , April 2015.
13. B. Wang , Yao Zheng, Wenjing Lou, Y. Thomas Hou, DDoS attack protection in the era of cloud computing and Software-Defined Networking, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, IEEE Communications Magazine, April 2015
14. R. Masoudi, Ali Ghaffari, Software defined networks: A survey, Department of Computer Engineering,Tabrizbranch, Islamic Azad University,Tabriz,Iran, Journal of Networkand Computer Applications 67(2016)1–25
15. A. Gonsalves, Prices Fall, Services Rise in Malware-as-a-Service Market, Mar. 2013. http://www.csoonline.com/article/2133045/malware-cybercrime/prices-fall-services-rise-in-malware-as-a-service-market.html
16. S. S. Silva, R. M. Silva, R. C. Pinto, and R. M. Salles, “Botnets: A survey,” Comput.Netw., vol. 57, no. 2, pp. 378–403, Feb. 2013.
17. C. Catalin, You Can Now Rent a Mirai Botnet of 400,000 Bots, Accessed on July, 09 2017 [Online]. Available at :https://www.bleepingcomputer.com/news/security/you-can- now-rent- a-mirai- botnet-of-400- 000-bots/
18. M. R. Haque, S. C. Tan, Z. Yusoff, L. C. Kwang, Motivation of DDoS Attack-Aware in Software Defined Networking Controller Placement, 2017, IEEE International Conference on Computer and Applications (ICCA), Doha, 2017, pp. 36-42.
19. SDXcentral, What is an OpenFlow Controller?, Accessed on 17 April 2018: https://www.sdxcentral.com/sdn/definitions/sdn-controllers/openflow-controller/

Malay Language Speech Recognition for Preschool Children using Hidden Markov Model (HMM) System Training

Marlyn Maseri and Mazlina Mamat

Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia
marlyn.maseri@gmail.com, mazlina@ums.edu.my

Abstract. This paper aims to discuss the implementation of Malay Language Speech Recognition (SR) using Hidden Markov Model (HMM) system training for Malay preschool children. The system is developed by implementing the architecture of HMM-based Recognizer with different feature extraction algorithm. The system is trained for 16 Malay words by collecting 704 speech samples (640 for training samples and 64 for testing samples). Data is collected from 20 preschool children aged between five to six years old in real time environments. The paper also describes the process flow to develop the architecture of the system. The experimental results show that the highest overall system performance is 94.86% - Train and 76.04% - Testing which is using MFCC (Mel-Frequency Cepstral Coefficient) with 39 extracted feature vectors (MFCC39).

Keywords: Malay Speech Recognition, Preschool Children, HMM, Malay Language, MFCC, LPC.

1 Introduction

Speech recognition (SR) is an important field of study and research and until now, studies on SR continued being explored due to varying nature of speech and languages. Recently, many researches have focused on developing SR to be utilized in learning and speech therapy system, and most of them are based on English language. In contrast to English SR, speech recognition specifically designed for Malay language are still limited [1]. Malay language has variations in certain pronunciations, even for the same word because of geographical boundaries and historical effects in different locations. Malay language belongs to the western branches of Austronesian languages (Malayo-Polynesian [2][3]) and it is adapted from the Arabic language [4]. Different from the sound systems of English language, Malay language is composed of [1]: (i)19 primary consonant that can be classified as stops (voiced and voiceless), fricatives (voiced and voiceless), affricates, nasals, lateral, and alveolar still, (ii)8 secondary consonants, (iii)6 vowels and (iv)3 diphthongs. Thus, developing SR targeting Malay language becomes more challenging and unpredictable.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_21

Previously, Tan et al. (2007) developed HMM-based Malay SR using 20 speech samples from 10 university students and implemented the SR system in their Malay Speech Therapy Assistance Tools (MSTAT) [5]. In addition, three studies have developed SR system but only for Malay vowel speech sound (‘*a*’,’*e*’,’*i*’,’*o*’,’*u*’). The system is developed using children samples (i.e., 360 children from 7 to 12 years [4] and 40 children from 7 to 10 years [6] respectively), and also 40 Malay undergraduate students [7]. The developed SR system is then implemented in their speech therapy system to be used by speech therapist to perform therapy on children with speaking difficulty and speech disorder [4][6] and also as assistive language tool for learners to pronounce Malay words properly and clearly [7].

However, even though many studies have been done in developing Malay SR, there has been little research on the applicability of Malay SR that specifically targeting preschool children [4][8]. SR system for preschooler is different from the adult as their speech is different in terms of absolute values and variability of acoustic and linguistic correlates[9]. Interestingly, studies have found out that in general, age has significant effect on formant frequencies whereby young Malay children has higher formant frequencies compared to adult [10] due to the length of their vocal tract [11]. Therefore, there are variations of the characteristics of children’s speech based on their age groups [12]. With this, it is necessary to specifically design Malay SR for preschooler by using their speech corpus instead of the adult speech corpus to match the characteristic of their speech. However, it is not easy to develop SR system for preschooler due to the complexity to build a speech corpus especially for resource constraint languages like Malay language [13]. With the limited research on Malay SR for preschooler, the development of Malay SR targeting preschool children is necessary so that they may experience the same benefits of SR in learning, speech therapy, entertainment and etc.

This paper describes the implementation of the Hidden Markov model (HMM) algorithm in Malay SR system for preschool children. For feature extraction, the system tested four different methods includes (i) LPC (Linear Predictive Coding) and MFCC (Mel-Frequency Cepstral Coefficient) with different feature dimension, (ii) MFCC-12, (iii) MFCC-26 and (iv) MFCC-39. The paper is organized as follows: Section 2 describes the methods used to develop the SR system includes data collection for speech corpus, feature extraction and HMM training. Section 3 shows the system performance and accuracy of the developed SR system. Finally, Section 4 concludes the paper and includes possible future work.

2 Methodology

The Malay SR implemented the HMM methods to train the system. The architecture of the Malay SR system follows the general HMM-based Recognizer as shown in Fig 1. The system is developed using Hidden Markov Model Toolkit (HTK)[14]. The development flow of the system is as shown in Fig. 2 which comprises of five main parts: Construction steps, Data preparation, Feature Extraction, HMM Training and Testing and Result Analysis.

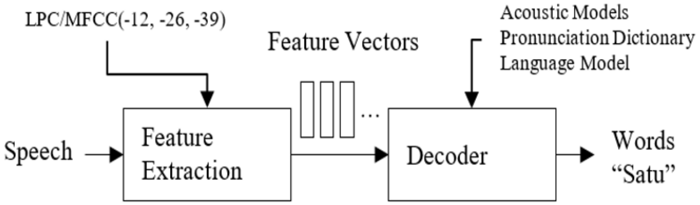
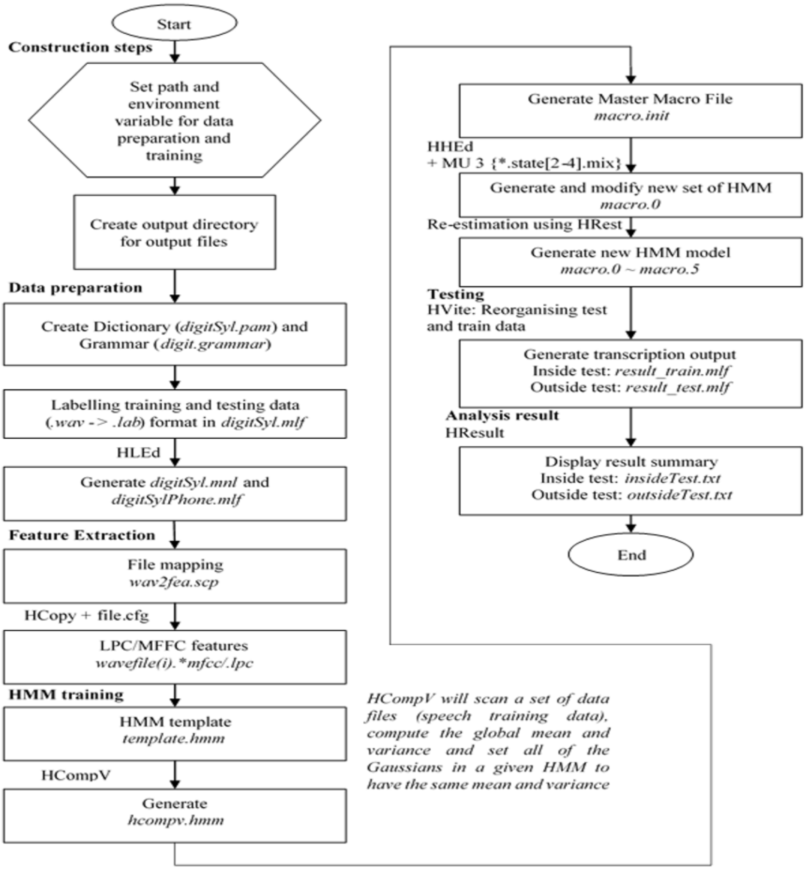


Fig. 1. Architecture of HMM-based Recognizer


Start
Construction steps
Set path and environment variable for data preparation and training
Create output directory for output files
Data preparation
Create Dictionary (*digitSyl.pam*) and Grammar (*digit.grammar*)
Labelling training and testing data (*.wav -> .lab*) format in *digitSyl.mlf*
HLEd
Generate *digitSyl.mnl* and *digitSylPhone.mlf*
Feature Extraction
File mapping *wav2fea.scp*
HCopy + file.cfg
LPC/MFFC features *wavefile(i).*mfcc/.lpc*
HMM training
HMM template *template.hmm*
HCompV
Generate *hcompv.hmm*
Generate Master Macro File *macro.init*
HHEd
+ MU 3 {*.state[2-4].mix}
Generate and modify new set of HMM *macro.0*
Re-estimation using HRest
Generate new HMM model *macro.0 ~ macro.5*
Testing
HVite: Reorganising test and train data
Generate transcription output
Inside test: *result_train.mlf*
Outside test: *result_test.mlf*
Analysis result
HResult
Display result summary
Inside test: *insideTest.txt*
Outside test: *outsideTest.txt*
End

HCompV will scan a set of data files (speech training data), compute the global mean and variance and set all of the Gaussians in a given HMM to have the same mean and variance

Fig. 2. Flow of the Speech Recognition Development Process

2.1 Speech Corpus

Ten Malay digits namely, Kosong(0), Satu(1), Dua(2), Tiga(3), Empat(4), Lima(5), Enam(6), Tujuh(7), Lapan(8), and Sembilan (9) and other six selected words (i.e., body parts) like Telinga, Mata, Tangan, Kaki, Mulut, and Hidung are used for our speech corpus. The selected words are words that easily distinguished and pronounced by preschool children aged between 5 to 6 years old. The word structure and phoneme sequence of each word are as shown in **Table 1**. In this corpus, all words are consisted of combination of syllables with two sounds, vowels (V) and consonants (C). There are four different syllable structures identified which are (i) V-2, (ii) CV-17, (iii) VC-2, (iv) CVC-8, (v) CCV-1, (vi) CCVC-1, (vii) CCVC-1 and (vii) CVCC-1. A total of 33 syllables are included in the corpus {/sa/, /tu/(2), /du/, /a/, /ti/, /ga/, /em/, /pat/, /li/(2), /ma/, /e/, /nam/, /juh/, /la/, /pan/, /sem/, /bil/, /an/, /ma/, /ta/(2), /hi/, /dung/, /mu/, /lut/, /ngan/, /ka/, /ki/, /te/, and /nga/}.

Table 1. Malay Words Speech Lexicon

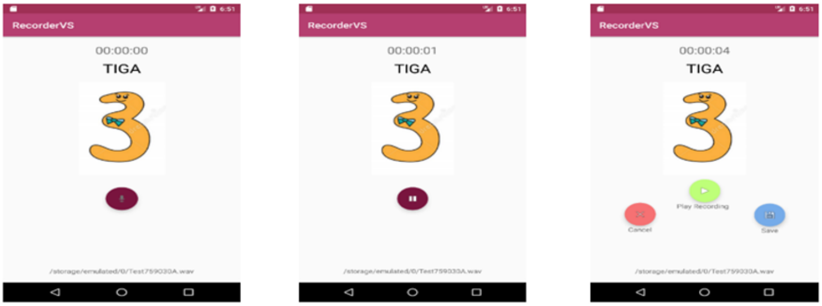
No	Word	Word in Malay	Phoneme Sequence	Word Structure	Syllables No
1	0	KOSONG	/ko song/	CV+CVCC	2
2	1	SATU	/sa tu/	CV+CV	2
3	2	DUA	/du a/	CV+V	1
4	3	TIGA	/ti ga/	CV+CV	2
5	4	EMPAT	/əm pat/	VC+CVC	2
6	5	LIMA	/li ma/	CV+CV	2
7	6	ENAM	/ə nam/	V+CVC	2
8	7	TUJUH	/tu juh/	CV+CVC	2
9	8	LAPAN	/la pan/	CV+CVC	2
10	9	SEMBILAN	/sem bi lan/	CVC+CVC+VC	3
11	Ear	TELINGA	/te li nga/	CV+CV+CCV	3
12	Eye	MATA	/ma ta/	CV+CV	2
13	Hand	TANGAN	/ta ngan/	CV+CCVC	2
14	Leg	KAKI	/ka ki/	CV+CV	2
15	Mouth	MULUT	/mu lut/	CV+CVC	2
16	Nose	HIDUNG	/hi dung/	CV+CVCC	2

2.2 Data Collection

Twenty preschool students were selected and they were aged between 5 to 6 years old. Therefore, a total of 704 speech samples (640 for training samples and 64 for testing samples) were recorded (two samples for each seventeen words) for every speaker. All speech samples were recorded using our own sound recording app, RecorderVS with the SGM-28 Omnidirectional Lavalier microphone. For each word, two samples were recorded (XXX0a.wav, XXX0b.wav, XXX:Name + Auto Generated Number + No). Each word to be pronounce was displayed in the app and the speaker required to pronounce the displayed word once at a time. The interface of the recording application is as shown in Fig. 3.

2.3 Feature Extraction

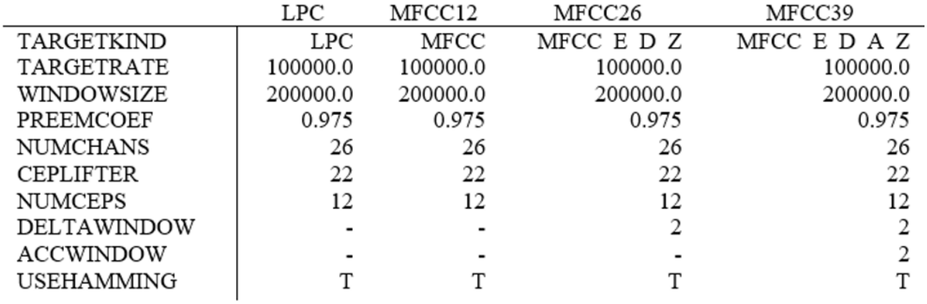
Features extraction in SR system involving the computation of a sequence of feature vectors in a more compact and robust form as representation of the speech signal before undergo the training process. In our system, the feature extraction is performed based on the following process:


RecorderVS
00:00:00
TIGA
/storage/emulated/0/Test759030A.wav
RecorderVS
00:00:01
TIGA
/storage/emulated/0/Test759030A.wav
RecorderVS
00:00:04
TIGA
Play Recording
Cancel
Save
/storage/emulated/0/Test759030A.wav

(i) Start Recording (iii) Recording (iii) End Recording

Fig. 3. The RecorderVS Interface

Pre-processing. Pre-processing consists of three steps which are pre-emphasis, framing and windowing. Pre-emphasis is performed by a first-order high-pass filter using a filter coefficient of ∝= 0.975 (*PREEMCOEF* value in Fig. 4).



	LPC	MFCC12	MFCC26	MFCC39
TARGETKIND	LPC	MFCC	MFCC_E_D_Z	MFCC_E_D_A_Z
TARGETRATE	100000.0	100000.0	100000.0	100000.0
WINDOWSIZE	200000.0	200000.0	200000.0	200000.0
PREEMCOEF	0.975	0.975	0.975	0.975
NUMCHANS	26	26	26	26
CEPLIFTER	22	22	22	22
NUMCEPS	12	12	12	12
DELTAWINDOW	-	-	2	2
ACCWINDOW	-	-	-	2
USEHAMMING	T	T	T	T

Fig. 4. Configuration for Feature Extraction

Commonly, the filter coefficient values between $0.9 \leq \propto \leq 1.0$. The pre-emphasis can be expressed as a difference in time domain as shown in Eq. 1.

$$Emphasis\ signal_t = Signal_t - \propto * Signal_{t-1} \tag{1}$$

After the pre-emphasis, the speech sound is turned into sharper with a smaller volume than the original speech signal. Next step is framing whereby the signal is split into short-time frames. For our system, we used frame shift of 10msec and frame length of 20msec (Frame shift; TARGETRATE and frame length; WINDOWSIZE values are as shown in Fig. 4). Window functions used is Hamming window. Hamming window is selected to avoid discontinuities. In Hamming window, the values of the signal are shrink toward zero at the boundaries. The equation for Hamming is as shown below (assuming a window that is L frames long):

$$hamming \qquad w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right), 0 \leq n \leq L-1 \\ 0, \qquad\qquad\qquad otherwise \end{cases} \tag{2}$$

Acoustic parameters. The final parameter vector layout for the extracted feature for each feature extraction are as shown in Fig. 5. For MFCC and LPC, only 12 basic coefficients are included. In MFCC26, there are 26 feature vectors which are 12 Mel

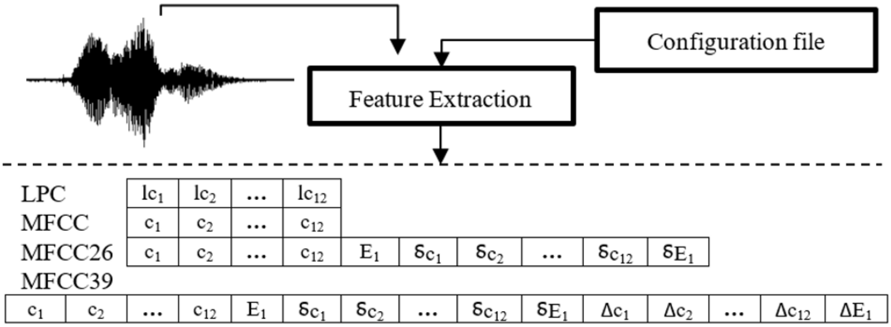
Cepstrum coefficients plus log energy and their delta (first order derivative) coefficients. The energy is computed as the log of the signal energy for speech samples $\{s_n, n = 1, N\}$:

$$E = \log \textstyle\sum_{n=1}^{N} s_n^2 \tag{3}$$

Next, the Delta coefficients (TARGETKIND: '_D') are computed using the following regression formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{4}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficient $(c_{t+\Theta})$ to $(c_{t-\Theta})$. The value of Θ is set using the configuration parameter, DELTAWINDOW (see Fig. 4**)**. MFCC39 is similar with MFCC26 but added second order regression coefficients, the Acceleration coefficients (TARGETKIND: '_A'). The formula is the same as Eq. 4. but the value of Θ is set using the configuration parameter ACCWINDOW**.**



LPC	lc_1	lc_2	…	lc_{12}															
MFCC	c_1	c_2	…	c_{12}															
MFCC26	c_1	c_2	…	c_{12}	E_1	δc_1	δc_2	…	δc_{12}	δE_1									
MFCC39																			
c_1	c_2	…	c_{12}	E_1	δc_1	δc_2	…	δc_{12}	δE_1	Δc_1	Δc_2	…	Δc_{12}	ΔE_1					

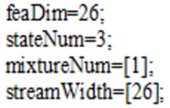
Fig. 5. Parameter Vector Layout for MFCC12, MFCC26, MFCC39

2.4 HMM Training

HMM is a double stochastic process, which consists of (i) underlying stochastic process that is not observable, and (ii) another set of stochastic processes that produce the sequence of observed symbols that is known. The underlying transition process that is hidden is a Markov process [15]. It is a probabilistic sequence model, which functions to assign a label to each unit (i.e. word, letter, morpheme) in a sequence and compute a probability distribution over possible sequences of labels before mapping it to the best label sequence. The probability distribution is represented by a mixture Gaussian density as shown in Eq. 5., where M_{js} is the number of mixture components in state j for stream s, c_{jsm} is the weight of m'th component and $\mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ.

$$b_j(o_t) = \textstyle\prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm})\right]^{\Upsilon_s} \tag{5}$$

To begin, the single set of single-Gaussian monophone HMMs is created where every mean and variance is identical. The monophones are then retrained by adding short-pause models and extending the silence model. The configuration for the HMM training is as shown in Fig. 7.



```
feaDim=26;
stateNum=3;
mixtureNum=[1];
streamWidth=[26];
```

Fig. 6. Configuration setting for HMM training

In our HMM training, the prototype model of the HMM template file (*template.hmm*) is initially created. The template is used to specify the model structure, which includes the number of states in an acoustic model, the number of streams in each state, and the number of Gaussian components in a stream. Next, the mean and variance of each Gaussian have been changed to the same values for all components and stored in new template (*hcompv.hmm*). These values are obtained via MLE (maximum likelihood estimate) based on all corpus. From the contents of *hcompv.hmm*, new prototype Master Macro File model (*macro.init*) which contains a copy for each of the required monophone HMMs is created. Next, re-estimation iteration is done to estimate the optimal values for the HMM parameters (transition probabilities, plus mean and variance vectors of each observation function). The re-estimation process is completed using the Baum-Welch method, which involves finding the probability of being in each state at each time frame using the Forward-Backward algorithm. In our system, re-estimation is repeated five times and each new HMM set is stored in a new directory (*macro.1*, *macro.2*, …, and *macro.5*). The final re-estimation created *macro.5* which stores the final set of initialized monophone HMMs.

3 Performance Analysis

In order to analyze the system performance and accuracy, we compare the machine transcription of the test utterances with the corresponding reference transcription files. The performance of SR system is evaluated by measuring the Accuracy (Eq. 6) and the Word Error Rate (Eq. 7), where N is the number of words in test set, D is the number of deletions, S is number of substitutions, H is the number of correct labels and *I* is the number of insertion.

$$\%acc = \frac{N-D-S-I}{N} \times 100\% = \frac{H-I}{N} \times 100\% \tag{6}$$

$$word\ error\ rate\ (WER) = \frac{S+I+D}{N} \times 100\% = 100 - \%acc \tag{7}$$

Training (Inside Test) and testing (Outside Test) is conducted to test the performance of the speech recognition. The result for Training and Testing are as shown in Table 2 and Table 3 respectively. Based on the result, MFCC with 39 feature vectors (MFCC39) has overall highest performance accuracy with lowest word error rate for both training (acc=94.86%, WER=4.90%) and testing (acc=76.04%, WER=23.80%). In contrast, LPC has the lowest performance accuracy with highest word error rate for both training (acc=57.98%, WER=42.00%) and testing (acc=19.99%, WER=76.40%). We could conclude that selection of feature extraction gives a huge impact on the performance of SR system. Based on the experiment, there are large difference in percentage between the highest (MFCC39) and lowest (LPC) performance accuracy which are 36.88% for training and 56.05% for testing. The system using LPC parameter has the lowest accuracy due to its linear computation nature [16]. Since the nature of human voice is nonlinear, LPC is not a good selection for feature extraction compared to MFCC. MFCC has better performance accuracy as it is designed based on the human peripheral auditory system which used an approximation to the Mel-frequency scale which is approximately linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz to follow the human auditory system which becomes less frequency-selective as frequency increases above 1 kHz [17].

Table 2. Performance of the SR – Training

	LPC		**MFCC12**		**MFCC26**		**MFCC39**	
Word	*acc*%	WER %	*acc*%	WER %	*acc*%	WER %	*acc*%	WER %
KOSONG	92.90	0.50	96.40	0.20	100.00	0.00	96.40	0.20
SATU	50.00	3.20	82.10	1.10	85.70	0.90	85.70	0.90
DUA	68.00	1.80	96.40	0.20	100.00	0.00	96.40	0.20
TIGA	32.10	4.30	82.10	1.10	85.70	0.90	89.30	0.70
EMPAT	69.20	1.80	85.70	0.90	85.70	0.90	96.40	0.20
LIMA	57.10	2.70	78.60	1.30	92.90	0.40	92.90	0.40
ENAM	57.10	2.50	82.10	1.10	96.40	0.20	96.40	0.20
TUJUH	42.90	3.70	96.30	0.20	100.00	0.00	96.40	0.20
LAPAN	50.00	3.00	71.40	1.80	96.40	0.20	92.90	0.40
SEMBILAN	70.40	1.80	92.90	0.40	100.00	0.00	100.00	0.00
TELINGA	71.40	1.80	71.40	1.80	89.30	0.70	96.40	0.20
MATA	28.60	4.60	70.40	1.80	92.90	0.40	89.30	0.70
TANGAN	53.60	3.00	71.40	1.80	92.90	0.40	92.90	0.40
KAKI	53.60	3.00	82.10	1.10	100.00	0.00	100.00	0.00
MULUT	59.30	2.50	82.10	1.10	100.00	0.00	96.40	0.20
HIDUNG	71.40	1.80	85.70	0.90	100.00	0.00	100.00	0.00
Total	**57.98**	**42.00**	**82.94**	**16.80**	**94.87**	**5.00**	**94.86**	**4.90**

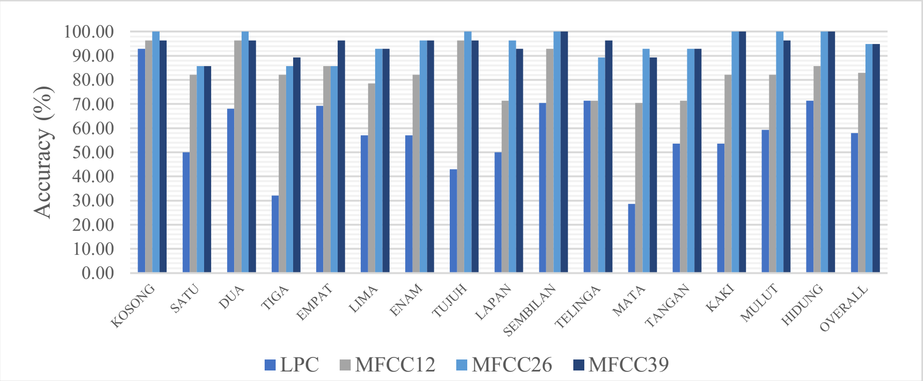


Fig. 7. Percent accuracy for Training (Inside Test)

Table 3. Performance of the SR – Testing

	LPC		**MFCC12**		**MFCC26**		**MFCC39**	
Word	*acc*%	WER %	*acc*%	WER %	*acc*%	WER %	*acc*%	WER %
KOSONG	50.00	3.20	83.30	1.10	83.30	1.00	100.00	0.00
SATU	25.00	4.80	58.30	2.60	41.70	3.60	58.30	2.60
DUA	16.70	5.40	41.70	3.70	75.00	1.60	83.30	1.00

TIGA	16.70	5.40	25.00	4.70	100.00	0.00	100.00	0.00
EMPAT	16.70	5.40	16.70	5.30	91.70	0.50	91.70	0.50
LIMA	36.70	3..8	16.70	5.30	41.70	3.60	50.00	3.10
ENAM	10.00	4.80	16.70	5.30	66.70	2.10	50.00	3.10
TUJUH	25.00	4.80	50.00	3.20	75.00	1.60	83.30	1.00
LAPAN	8.30	5.90	9.10	5.30	50.00	3.10	66.70	2.10
SEMBILAN	8.30	5.90	58.30	2.60	83.30	1.00	75.00	1.60
TELINGA	36.40	3.80	25.00	4.70	58.30	2.60	66.70	2.10
MATA	0.00	6.50	27.30	4.20	66.70	2.10	66.70	2.10
TANGAN	8.30	5.90	33.30	4.20	83.30	1.00	83.30	1.00
KAKI	0.00	6.50	50.00	3.20	91.70	0.50	83.30	1.00
MULUT	41.70	3.80	33.30	4.20	83.30	1.00	83.30	1.00
HIDUNG	20.00	4.30	25.00	4.70	66.70	2.10	75.00	1.60
Total	**19.99**	**76.40**	**35.61**	**64.30**	**72.40**	**27.40**	**76.04**	**23.80**

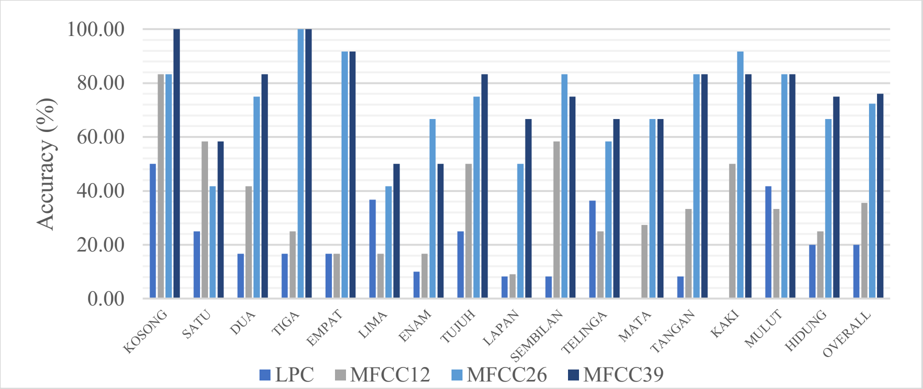


Fig. 8. Percent accuracy for Testing (Outside Test)

4 Conclusion

This paper addressed the principle of HMM algorithm with different feature extraction for SR system. Techniques used to implement the algorithms in the system are described in details in the methodology section. From the results, it can be concluded that HMM with MFCC as feature extraction can recognized speech signal better than LPC. The development of a Malay SR using an HMM-based speech synthesis system for Malay preschool children is important because Malay is an under-resourced language and research on SR specifically for Malay speaking preschool children is limited. In our system, we only used 16 selected words with 704 speech samples. It is a challenging procedure to collect a speech corpus from preschool children (5 to 6 years old) as some of them could not pronounce the selected words correctly and therefore, they are excluded from our corpus. For example, they pronounced the word "HIDUNG" as "/i/ dung/" with silent 'h' word instead of "/hi/ /dung/", and also mispronounce the long word with three syllables like the word SEMBILAN; /sem/ /bi/ /lan/ - CVC+CVC+VC). Future study should be done by targeting more Malay words with large number of speech samples to improve performance of the existing speech recognition.

5 References

1. Ong, H.F., Ahmad, A.M.: Malay Language Speech Recogniser with HHM Model and ANN. International Journal of Information and Education Technology 1(2), 114–119 (2011).
2. Fook, C.Y., et al.: Malay speech recognition in normal and noise condition. In: 8th IEEE International Colloquium on Signal Processing and its Applications, pp. 409–412 (2012).
3. Mustafa, M.B., et al.: Context-dependent labels for an HMM-based speech synthesis system for Malay HMM-based speech synthesis system for Malay. In: International Conference O-COCOSDA held jointly with CASLRE, pp. 1–4. IEEE, Gurgaon (2013).
4. Zourmand, A., Nong, T.H. Intelligent Malay Speech Therapy System. In: IEEE Conference on Systems, Process and Control (ICSPC), pp. 111–116. IEEE, Malacca (2017).
5. Tan, T-S., et al.: Application of Malay speech technology in Malay Speech Therapy Assistance Tools. In: 2007 International Conference on Intelligent and Advanced Systems, pp. 330–334. IEEE, Kuala Lumpur (2007).
6. Hua, N. T., Yunus, J.: Speaker-independent Malay vowel recognition of children using multi-layer perceptron. IEEE Region 10 Conference TENCON, Vol. 1, pp. 68–71 (2004).
7. Mohd Yusof, SA.: Development of Malay Word Pronunciation Application using Vowel Recognition. International Journal of u- and e- Service, Science and Technology 9(1), 221-234 (2016).
8. Azmi, MYS.: Malay Word Pronunciation Test Application for Pre-School Children. Int Journal Of Interactive Digital Media 4(2), 2289–4098 (2016)
9. Yildirim, S., et al.: Acoustic analysis of preschool children's speech. In: Proc. 15th ICPhS, pp. 949–952 (2003).
10. Ting, HN., et al.: Formant frequencies of Malay vowels produced by Malay children aged between 7 and 12 years. Journal of Voice 26(5), 664.e1-664.e6 (2012).
11. Zourmand, A., et al.: The Effect of Age on Formant Frequencies of Malay Children between 7-12. In: Third International Conference on Intelligent Systems Modelling and Simulation. IEEE, pp. 379–382 (2012).
12. Giuliani, D., Gerosa, M.: Investigating recognition of children's speech. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP '03), pp. II-137-40 (2003).
13. Rahman, FD., et al.: Automatic speech recognition system for Malay speaking children. In: 2014 Third ICT International Student Project Conference (ICT-ISPC), pp. 79–82. IEEE, Nakhon Pathom (2014).
14. HTK Speech Recognition, http://htk.eng.cam.ac.uk/, last accessed 2018/04/30.
15. Gales, M., Young, S.: The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing 1(3), 195–304 (2007).
16. Dave, N.: Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition. International Journal for Advance Research In Engineering And Technology 1(5), 1–5 (2013).
17. Noushad, A., et al.: A Study on Different Music Genre Classification Methods. International Journal of Computer Science and Mobile Applications 6(2), 131–138 (2018).

A Formal Model of Multi-agent System for University Course Timetabling Problems

Kuan Yik Junn, Joe Henry Obit, Rayner Alfred and Jetol Bolongkikit

Knowledge Technology Research Unit, Universiti Malaysia Sabah,
Labuan International Campus, Malaysia
kyj_anderson@hotmail.com, joehenryobit@gmail.com,
ralfred@ums.edu.my, jetol@ums.edu.my

Abstract. This paper describes a general framework of Multi-agent system which incorporates the hyper-heuristics search methodology with both Great Deluge and Simulated Annealing acceptance criteria respectively. There are three types of agents introduce in the framework which involve the communication between heuristic agents, cooperative agents and mediator agent. The common goal for each agent is to improve the quality of course timetabling solutions until the best solution is found when the termination condition meets. A preliminary experiment have been conducted towards this approach in university course timetabling problem and the results shows the framework is able to increase the quality of existing solution compared with other meta-heuristics which have been studied in the previous researches.

Keywords: Cooperative Searches, Multi-agent Systems, University Course Timetabling Problems.

1 Introduction

Multi-agent systems (MAS) are evolution sub-field of artificial intelligence which interacts with society of agents in order to achieve common objective. This field has been studied since 1980s and gained attention worldwide after ten years [13, 17]. Multi-agent have become increasingly important in various aspects by highlighting the issues of interaction and distributed intelligence which represent a new method of analyzing, developing, designing and implementation of advance software systems in computer science. In MAS, agent is an entity which acts in an environment, although there are various definitions which defined the term in the literatures [1]. Therefore, a population of autonomous agent working together in an environment in order to reach common objective formed MAS [2].

The application of MAS in university course timetabling problem (UCTP) is still a growing trend in optimization problems. Local searches or heuristics are the most common techniques applied in this scheduling problem but the weakness is it often lead the solution stuck in local optima. From there, many metaheuristics searches such as Great Deluge, Simulated Annealing, Tabu Search as well as Genetic Algorithm have been introduced and investigated in this domain. Metaheuristics are considered

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_22

to be domain dependent due to parameters in each algorithm which need to adjust to suits the particular domain [10]. Hence, the drawback motivates the researchers to study on hyper-heuristics and MAS. In this research, a real-world problem on UCTP is studied using two datasets obtain from the institution.

This paper highlight the section as follow: Section 2 describes about related works done on multi-agent and UCTP while Section 3 discusses about the problem background of this research. Section 4 presents the formal mathematical model based on the domain constraints while experimental results are discussed in Section 5 and the last section summarize the research of this paper.

2 Related Works

Some of the recent work involved the implementation of Great Deluge [9] and Simulated Annealing [4] potentially producing good solutions in UCTP. The implementation of metaheuristics improves the quality of solution much better in some of the domain which has been studied previously [8]. Ines [15] proposed a Multi-agent based hyper-heuristic algorithm for Winner Determination Problem which explores a set of cooperating agents in order to choose the most suitable method using learning techniques. The proposed algorithm is specialized for local search methods and evolutionary strategies where different types of agents namely mediator agent, local search agents, perturbation agent and recombination agents are searching to improve the solution, which show that it performs well on benchmark instances.

The cooperative search in multi-agent involves several types of agent with different behaviors. The main goals are to reduce the computation time needed to solve problem instances and enhancing the robustness of search in the same computation time. The cooperative hyper-heuristic search framework from the literature [14] proposed a generic model with two types of agents such as heuristic agents and cooperative hyper-heuristic agents. The heuristic agent perform local search to improve the local solution and cooperate with cooperative hyper-heuristic agent by exchanging the local best solutions while cooperative hyper-heuristic agent manages the cooperation between heuristic agent, select the heuristic and the acceptance move.

Another similar study using Particle Swarm Optimization in cooperative distributed hyper-heuristic for course timetabling problem [7] is able to improve the solution quality for large instance, find new best solution for all five medium and five small in benchmark datasets obtained from the literature [16] with the proposed Asynchronous Cooperative Distribute Low-level heuristics (ACDLLHs) algorithm. In different approach to generate initial timetable using multi-agent, several hybrid heuristics which combine with graph coloring in different ways proven to be potentially giving promising results [12].

Inspired by the characteristics of agent in terms of communication and complex behavior, this research presents a general framework of MAS adopted from some of the past researches [14] which involves two levels of communication. A hyper-heuristic search is adopted using reinforcement learning modified from the model [11] and applied in the proposed framework.

3 Problem Background

The focus in this research is on Universiti Malaysia Sabah Labuan International Campus (UMSLIC). There are two types of constraints which need to be tackle: hard constraints and soft constraints. In this paper, hard constraints are important and need to be satisfied while soft constraint should be satisfied as much as possible. Two sets of datasets are gathered from Academic Service Division (ASD) of UMSLIC.
The hard constraints for the UMSLIC course timetabling are as follow:

1. (Hc_1) Student should not attend more than one class at the same time.
2. (Hc_2) The capacity of a room must be larger than the size of the course enrolled by the students.
3. (Hc_3) There must be not more than one course assigned to the same room at the same time.
4. (Hc_4) For each different type of course must satisfied their respective predefined timeslot stated by the ASD department.

The soft constraints for the UMSLIC course timetabling are as follow:

1. (Sc_1) Student should not attend more than two classes consecutively.
2. (Sc_2) The remaining empty seats in the difference between the capacity of room and size of course enrollment should be minimized as much as possible.

The UCTP is different from other timetabling problem based on their respective institution due to different number of constraints [8]. For example, the datasets summary for UMSLIC is stated below:

Table 1. Summary of datasets in UMSLIC.

Datasets	No. of students	No. of courses	No. of timeslots	No. of rooms
Sem 2 s2014/2015	2248	112	35	18
Sem 1 s2015/2016	2371	125	35	18

4 Mathematical Model of Problem Instance

In this section, the mathematical model of UMSLIC UCTP is defined. The different types of courses available in UMSLIC are as follow: main courses; knowledge promotion courses; language subjects and co-curricular activities. Let C be the total number of courses and C_1 denote as total number of main courses, C_2 denote as total number of knowledge enhancement courses, C_3 denote as total number of language subject and C_4 denote as total number of co-curriculum activities. Thus, the total number of courses is calculated as follow: $C = C_1 + C_2 + C_3 + C_4$.

As stated previously, the main objective is to satisfy all the hard constraints, Hc = Hc_1, Hc_2, Hc_3, Hc_4 and minimize the violation of soft constraints as much as possible,

Sc_1, Sc_2 subject to Hc = 0. *F* can be denoted as the cost function, which can be calculated as follow:

$$F = Hc + Sc_1 + Sc_2 \tag{1}$$

In most research on UCTP, usually penalty is assigned to infinity value for every violation of hard constraints. In this research, the penalty for hard constraint is assigned with the weight of 100,000. The reason is given the worst case of violation of soft constraints will definitely not exceed 100,000. If the cost function exceeds that value, it means there is violation of hard constraints. There are 4 hard constraints which denote as ($\lambda_1\ until\ \lambda_4$) while λ_5 and λ_6 are both soft constraints. The penalty for every weight is summarized in Table 2:

Table 2. Penalty assignation of constraint violations.

Weight	Penalty	Description
$\lambda_1\ to\ \lambda_4$	100,000	Hard constraints, *Hc*
λ_5	5	Soft constraint, Sc_1
λ_6	2	Soft constraints, Sc_2

The mathematical models representation for both hard constraints and soft constraints are presented as follow. For hard constraint 1 (Hc_1) shown in equation (2), the penalty given if student attends more than one class at a time is given by

$$H_{C_1} = \lambda_1 \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} CC_{ij} \tag{2}$$

where *C* is the total number of courses, λ_1 is the weight, CC_{ij} is the number of students who have at least two classes i and j at the same time.

For hard constraint 2 (Hc_2) shown in equation (3), the penalty given if the capacity of room is smaller than the size of course enrolled by student is given by

$$H_{C_2} = \lambda_2 \sum_{i=1}^{C} \sum_{j=1}^{R} CR_{ij} \tag{3}$$

where *C* is the number of courses, *R* is the number of rooms, λ_2 is the weight, CR_{ij} is the number of rooms which the course *i* larger than the capacity of that room j.

For the hard constraint 3 (Hc_3) in equation (4), the penalty given if more than one course placed together at the same room at the same time is given by

$$H_{C_3} = \lambda_3 \sum_{i=1}^{R} \sum_{j=1}^{T} RT_{ij} \tag{4}$$

where *R* is the number of rooms, *T* is the number of timeslots λ_3 is the weight, RT_{ij} is the number of courses which have different lectures, room *i* and timeslot j at the same time and classroom.

For the hard constraint 4 (Hc_4) shown in equation (5), the penalty given if the specific type of courses assigned other than its predefined timeslot is given by

$$H_{C_4} = \lambda_4 \left(\sum_{i=1}^{c_1} \sum_{t=1}^{T} M_{it} + \sum_{j=1}^{c_2} \sum_{t=1}^{T} P_{jt} + \sum_{k=1}^{c_3} \sum_{t=1}^{T} L_{kt} + \sum_{l=1}^{c_4} \sum_{t=1}^{T} C_{lt} \right) \tag{5}$$

where c_1, c_2, c_3, c_4 are the total number of main courses, knowledge enhancement courses, language courses, and co-curricular courses respectively, λ_4 is the weight, T is the number of timeslot and M_{it}, P_{jt}, L_{kt} and C_{lt} are violation of the particular course *i, j, k, l* to be assigned in timeslot *t*.

Meanwhile, in terms of soft constraint 1 (Sc_1) shown in equation (6), penalty given if student attends more than two classes consecutively is given by

$$S_{C_1} = \lambda_5 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij} \tag{6}$$

where n is the number of class activities, λ_5 is the weight, C_{ij} is the consecutive classes attend by student of course i and j.

For soft constraint 2 (Sc_2) shown in equation (7), the penalty given if there are remaining spaces in the specific room of a particular timeslot is given by

$$S_{C_2} = \lambda_6 \sum_{i=1}^{n} \sum_{j=1}^{n} C_{i-j} \tag{7}$$

where n is the number of activities, λ_6 is the weight, C_{i-j} is the difference between the size of room i and course enrolment of course j.

5 Multi-agent framework

The Multi-agent framework is developed using Java Agent Development (JADE). The implementation of the framework is believed to be potential in producing good solutions. MAS is known to be an intelligent distributed approach which is well suited for applications which are changeable, complex, distributed and modular [14]. The agents cooperate in order to find a globally good schedule which can effectively response to real-time interference and optimize the original goal. Besides that, MAS is capable to maintain its functionality although there might be failure in an individual agent, which will not affect the entire system [13]. The proposed MAS in this research focus on general domain which covers the problem in university course timetabling. This framework consists of several agents which communicate with each other in order to improve the quality of the solution. There are three types of agents proposed:

- Heuristic agents – each of these agents is represented as single low-level heuristic which perform local search.
- Cooperative agents – the cooperative agent is composed of meta-heuristic algorithm which select low-level heuristic and performed improvement based on initial solution carried out by heuristic agents whereby two main phases are considered: selection strategy and acceptance criteria.
- Mediator agent – the role of mediator agent is accepting solutions from the cooperative agents and performing improvement functions using genetic algorithm. It will update the global best solution once it receives the solutions from cooperative agents.

There are two layers in the proposed framework. The first layer consists of communication between heuristic agents and cooperative agents. Commonly, initial solution is created with local search or heuristics. Some of the examples are using initialization heuristics and Tabu Search [6]. Other initialization heuristics and techniques can also be found in the literatures [3]. In this layer, the initial solution is constructed using heuristic agents. The heuristic agents may perform different types of local search and tries to produce a feasible solution. Once a feasible solution is produced, the heuristic agents will pass the solutions to cooperative agents for improvement. From there the cooperative agents perform hyper-heuristics search using GD and SA acceptance criteria. In selection strategy, an adaptive learning is implemented using Modified Choice Function (MCF) of static memory length as implemented from the literature [10]. In the MCF, reward will be given to the heuristic agent if it manages to improve the existing solution. Else no reward will be given. Meanwhile, the acceptance criteria are based on acceptance rate from both GD and SA which accept the local search movement for next iterative improvement. The continuity from the first layer of the framework updates the global best solution in the second layer of the framework. Hence, cooperative agents improve the quality of solution and forwarded it to centralized agent which is known as mediator agent. The mediator agent provides specific role and perform several tasks at a time. In this matter, there will be continuous interaction between the cooperative agents and mediator agent. The roles of cooperative agents in the second layer of the framework perform several tasks which are request new solution and improve it.

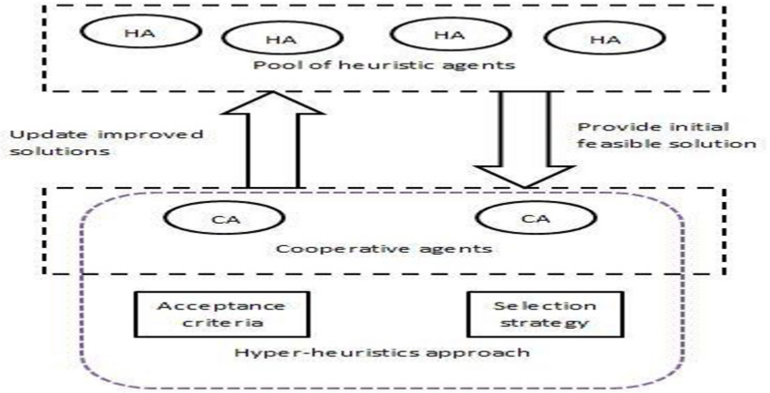


Fig. 1. First layer of Multi-agent framework

Each cooperative agent improves the solution and sends the solutions to mediator agent whereby mediator agent incorporates Genetic Algorithm search improvement. The mediator agent in this framework performs several tasks:

- Receives initial solutions from cooperative agents and allocate it into the pool.
- Provide existing solution from the pool of solutions to cooperative agents.

- Perform search improvements from the pool of solutions using genetic operators.
- Control the quality and update the global best solution for each solution it received.

Table 3. Specification of agent's knowledge and responsibilities

Agent type	Acquaintances	Responsibilities
Heuristic	Cooperative	1. Search for improving solutions. 2. Sends the solution to cooperative agents.
Cooperative	Heuristic, Mediator	1. Accepts all the local best solutions from heuristic agent. 2. Acceptance criteria are based on the acceptance probability from Simulated Annealing and Great Deluge. 3. Performing improving solutions for every iterations and send its current local best solution to the mediator agent as soon as the termination condition meets.
Mediator	Cooperative	1. Accepts initial solutions from cooperative agents to form a pool of solutions. 2. Accepts the local best solution from cooperative agents and select the overall local best solution which will be stored in pool of best solution. 3. Update the global best solution if there is better solution in the pool of best solution. 4. Perform genetic operators after the pool of best solution has reached maximum limit and the global best solution will be updated iteratively if there is improvement.

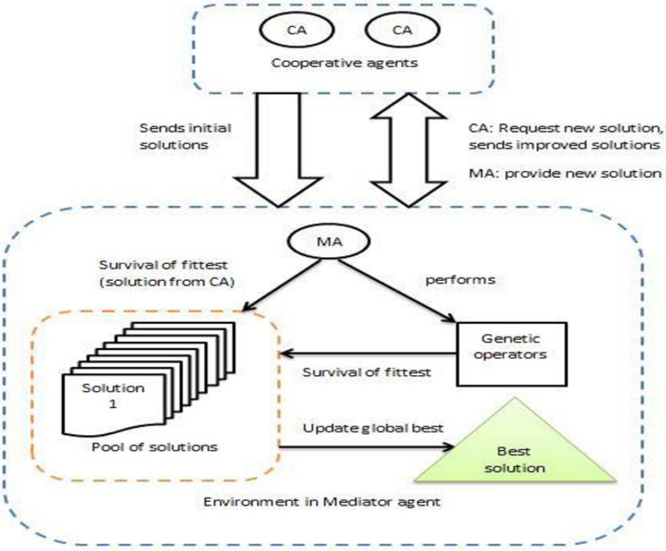


Fig. 2. Second layer of Multi-agent framework

The communication between agents in this stage shares the common goal which is to produce better quality of solutions. Figure 2 shows the framework in the second layer which involves communication between cooperative agents and a mediator agent. The environment state in the system is the search space of the solution while the state of action performed by the agents will indirectly affect the environment state whereby the landscape of search space will change based on the scale of influence. Any working environment might involve the presence of active agent to operate based on the internal state. The internal state of an agent is depending on the active behavior that agent may perform and also the characteristics of the algorithm. Hence, the use of JADE platform provides ontology for the agent to communicate in complex environment. The environment will not change if there is no action due to the absence of agent. The agents' knowledges and their responsibilities are summarized in Table 3.

6 Preliminary Results

The proposed framework is developed in order to determine the effectiveness of the model approach in JADE platform. A preliminary experiment for this framework is conducted towards educational timetabling domain. There are 2 tests experimented for preliminary results. Each test has been performed with 50 runs and the average cost function or penalty is recorded. The calculation of cost function is based on the violation of constraints in a specific domain. The lower the cost function, the better the quality of solution. The first test is to measure the number of agents against the performance. The results are recorded in every 5 minutes as shown in Figure 3.

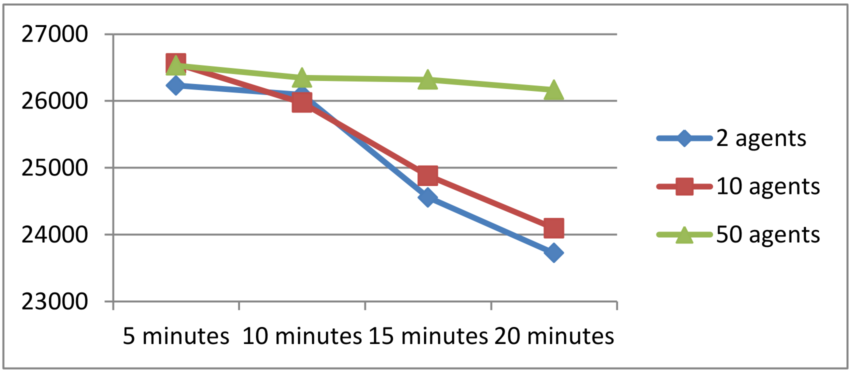


Fig. 3. Cost function against time based on number of agents

The results shown in Figure 3 indicate that increasing the number of agents in the platform might yield to poorer performance. The reason could be following factors:

- Increasing the number of cooperative agents involves more messages passing to mediator agent at a time, hence increasing the time for mediator agent to process the message.
- Distribution of solutions by mediator agent in every period might indirectly affect the message received by cooperative agent pending if it's still improving the cur-

rent solution. Hence, the messages will be accumulated in queue while the cooperative agents are still processing the current message.

Potentially the performance of multiple agents working in an environment can be improved by requesting of solution from the cooperative agents instead of distribution from the mediator agent. In the second experiment, a comparison of performance is tested between the proposed multi-agent frameworks with other metaheuristics search strategy which have been investigated in the previous researches [5]. In the second experiment, the approaches are performed on the same dataset of university course timetabling problem. The first dataset is dataset from institution during semester 2 sessions 2014/2015 and the second dataset is semester 2 sessions 2015/2016. In the first dataset, the average cost function performed by MA is lower than both GD and SA with given the same computation time. In second dataset, the average cost function produced is slightly lower than the other metaheuristics. Although the improvement done by SA or GD is less than of MA, the results showed significant improvement for both semesters. The average percentage improvement shown in Table 5 indicates the improvement from initial solution. The MA manages to improve the quality of solution with more than 30 percent for both semesters, which is much better than the other meta-heuristics. The overall results proved MA performs better than the other meta-heuristics in terms of average cost function.

Table 4. Experimental results between Multi-agent and metaheuristics.

Experiment	Semester 2 Session 2014/2015			Semester 1 Session 2015/2016		
Algorithm	MA	GD	SA	MA	GD	SA
Average cost	27192	29081	30139	28830	29400	30475
Average time	300.05	301.47	301.37	300.06	301.01	301.21
Improve (%)	33.56	25.87	23.62	30.33	25.44	22.41

7 Conclusion

In this research, a general cooperative search using multi-agent framework is proposed. Nonetheless, various multi-agent frameworks have already been proposed in the literatures. The proposed framework is considered to be new in this domain as mostly the implementation of MAS are applied towards transportation, networking, business analytic, management, strategic planning, etc. The proposed framework is composed of two levels and the communications between different types of agents in each level are explained in Section 5. The results have showed that cooperative approaches with multi-agent are more effective in improving the solution compare with normal meta-heuristics.

For future research, different problem domain will be used in order to produce generic framework. This may involves several important factors which need to be study such as the communication and coordination between agents. More algorithms can be implements into the framework such as Tabu Search to select the promising heuristic for improvements and the nature of Ant algorithm which can be suited into the behavior of an agent in order to cooperate with other agents.

References

1. Ferber, J.: Multi-agent systems: an introduction to distributed artificial intelligence, vol. 1. Reading: Addison-Wesley (1999).
2. Jennings, N. R., Sycara, K., & Wooldridge, M.: A roadmap of agent research and development. Autonomous agents and multi-agent systems, 1(1), 7-38 (1998).
3. Junn, K. Y., Obit, J. H., & Alfred, R.: A Constraint Programming Approach to Solving University Course Timetabling Problem (UCTP). Advanced Science Letters, 23(11), 11023-11026 (2017).
4. Junn, K. Y., Obit, J. H., & Alfred, R.: Comparison of Simulated Annealing and Great Deluge Algorithms for University Course Timetabling Problems (UCTP). Advanced Science Letters, 23(11), 11413-11417 (2017).
5. Junn, K. Y., Obit, J. H., & Alfred, R.: The Study of Genetic Algorithm Approach to Solving University Course Timetabling Problem. In International Conference on Computational Science and Technology, pp. 454-463. Springer, Singapore (2017).
6. Landa-Silva, D., & Obit, J. H.: Comparing Hybrid Constructive Heuristics for University Course Timetabling (2011).
7. Obit, J. H., Alfred, R., & Abdalla, M. H.: A PSO Inspired Asynchronous Cooperative Distributed Hyper-Heuristic for Course Timetabling Problems. Advanced Science Letters, 23(11), 11016-11022 (2017).
8. Obit, J. H., Junn, K. Y., & Alfred, R.: A Performance Comparison of Metaheuristics Search for University Course Timetabling Problems. Advanced Science Letters, 23(11), 11012-11015 (2017).
9. Obit, J. H., Junn, K. Y., & Alfred, R.: Performance Comparison of Linear and Non-Linear Great Deluge Algorithms in Solving University Course Timetabling Problems. Advanced Science Letters, 23(11), 11027-11030 (2017).
10. Obit, J. H., Landa-Silva, D., Sevaux, M., & Ouelhadj, D.: Non-linear great deluge with reinforcement learning for university course timetabling. Metaheuristics–Intelligent Decision Making, Series Operations Research/Computer Science Interfaces, 1-19. Springer (2011).
11. Obit, J. H., Landa-Silva, D., Ouelhadj, D. & Sevaux, M.: Non-linear great deluge with learning mechanism for solving the course timetabling problem. In: 8th Metaheuristics International Conference. MIC (2009).
12. Obit, J. H., Landa-Silva, D., Ouelhadj, D., Vun, T. K., & Alfred, R.: Designing a multi-agent approach system for distributed course timetabling. In Hybrid Intelligent Systems (HIS), 2011 11th International Conference on IEEE, pp. 103-108. (2011).
13. Oliveira, E., Fischer, K., & Stepankova, O.: Multi-agent systems: which research for which applications. Robotics and Autonomous Systems, 91-106 (1998).
14. Ouelhadj, D., & Petrovic, S.: A cooperative hyper-heuristic search framework. Journal of Heuristics, 16(6), 835-857 (2010).
15. Sghir, I., Jaafar, I. B., & Ghédira, K.: A Multi-Agent based Hyper-Heuristic Algorithm for the Winner Determination Problem. Procedia Computer Science, 112, 117-126 (2017).
16. Socha, K., Knowles, J., & Samples, M.: A Max-Min Ant System for the University Course Timetabling Problem. In Ant Algorithms: Proceedings of the Third International Workshop (ANTS), LNCS, vol. 2463, pp. 1-13. Springer (2002).

17. Wooldridge, M., Fisher, M., Huget, M. P., & Parsons, S.: Model checking multi-agent systems with MABLE. In Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2, ACM, pp. 952-959. (2002).

An Investigation towards Hostel Space Allocation Problem with Stochastic Algorithms

Joe Henry Obit, Kuan Yik Junn, Rayner Alfred, Jetol Bolongkikit and Ong Yan Sheng

Knowledge Technology Research Unit, Universiti Malaysia Sabah,
Labuan International Campus, Malaysia
joehenryobit@gmail.com, kyj_anderson@hotmail.com,
ralfred@ums.edu.my, jetol@ums.edu.my, yan.s@hotmail.com

Abstract. This research presents the study of stochastic algorithms in one of the limited study in Space Allocation Problem. The domain involves the allocation of students into the available rooms which is known as Hostel Space Allocation Problem. The problem background of this domain which related with hard constraints and soft constraints are discussed and the formal mathematical models of constraints in Universiti Malaysia Sabah Labuan International Campus are presented. The construction of initial solution is handled by Constraint Programming algorithm. Two algorithms mainly Great Deluge with linear and non-linear decay rate and Simulated Annealing with linear reduction are proposed to improve the quality of solution. The experimental results show that Simulated Annealing with linear reduction temperature performs well in this domain.

Keywords: Great Deluge, Simulated Annealing, Hostel Space Allocation Problem

1 Introduction

Hostel Space Allocation Problem (HSAP) is a new domain of complex combinatorial optimization problem as the research studied on this problem is limited. The availability and effective management of hostel space will affect the student's focus and academic performance, indirectly contributes towards the graduation success rate of an institution. Due to increasing numbers of student admission into higher learning institution worldwide, this issue has grown concern among institution and gained attention among researchers to study in this area. The HSAP is similar to scheduling problems in other domain such as University Course Timetabling Problem [7, 8, 9] (UCTP) and examination timetabling problem [10]. HSAP in higher institution can be defined as allocation of events (such as students) into available room in order to satisfy the list of constraints [1]. Generally, the need of institution administrator to manage the spaces for students to live is important which deal with limited number of resources among competitive and demanding number of students who apply to hostel in the institution. Therefore, it is required to think of an effective allocation techniques for limited resources. In such situation, the algorithm can be considered effective when all the demanding entities such as student are given very limited space allocation but satisfying

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_23

all the constraints. The use of suitable optimization strategy [4] in automated system is believed to minimize the time and effort for scheduling problem. Heuristic approach may search for optimal solutions within reasonable computational time but the solution tend to stuck in local optima and thus may not be effective for domain specific knowledge. Hence, meta-heuristic searches methodologies are proposed using both Great Deluge and Simulated Annealing algorithm in order to solve the HSAP domain.

The content of this paper is presented as follow: Section 2 discussed on the related works in HSAP and the meta-heuristics applied while Section 3 described on the problem background of this research. A formulation mathematical model on HSAP of this research is presented on Section 4 and the experimental results are discussed in Section 5. The last section concludes the content of this research.

2 Related Works

In general HSAP is categorized under space allocation problems (SAP) which is more generic domain-specific and gained attention among researchers to study on meta-heuristics [13]. In [1], space allocation is assigned manually in high institution, commonly in developing countries which required a lot of time to schedule. Several domain related to scheduling have been studied in the literature such as shelf space allocation [2], timetabling [14, 15, 16], and office space allocation [12]. The continuous growing number of applicants for limited hostel space and facilities need an effective optimization strategy in HSAP. Hence, this domain can be considered as the modification of knapsack problem [19].

Various techniques are used to tackle real-world problem in SAP such as mixed-integer programming [6] and linear programming [3]. In NP-hard combinatorial optimization problems, the algorithms used in SAP is expectedly to has exponential time complexity, whereby solving a large instance is definitely computational expensive. Therefore, a proper technique which suits well for the particular domain needs to be studied. The use of Simulated Annealing and other meta-heuristics from [2] in real-world shelf space allocation problem is introduced due to constraints such as limited shelf space and massive amount of products to be allocated for viewing. Apart from that, non-linear Great Deluge with reinforcement learning [17] also had been adopted in order to tackle course timetabling problem. In similar studies on SAP, Landa-Silva [12] conducted research on Tabu Search and Genetic Algorithm for office allocation problem which works well in that domain. A hierarchical-based heuristics solution is introduced in [1] for SAP which applied different inter-related heuristics at multi-level stage procedure in order to satisfy provided set of constraints. Regarding to that, Genetic Algorithm is implemented to guide the allocation to lowest level (best solution). The results show the strategy applied works efficiently to HSAP domain. In another similar research on HSAP, a multi-level Genetic Algorithm is applied on multi-stage real-world allocation problem occur in Nigerian universities [1]. The mete-heuristic is able to produce promising results within given computational time.

In this research, the Great Deluge and Simulated Annealing algorithms are proposed and compared the performance between them in order to identify the more appropriate and efficient strategy for this problem. To date, the application of metaheuristics to HSAP had only been done by Adewumi [1] using Genetic Algorithm approach, in which there is no existing datasets of HSAP for benchmarking purpose. However, the operators in Genetic Algorithm in [1] are computationally expensive as the process required longer time to perform the searches although it can produced good solutions. Hence, stochastic strategies are adopted in this research which can potentially produce good results in reasonable computational time.

3 Hostel Space Allocation Problem in Institution

The domain studied in this research focus on Universiti Malaysia Sabah Labuan International Campus (UMSLIC). The datasets are obtained from the Student Affair Department (HEP) of UMSLIC whereby hard constraints and soft constraints are identified from conducting interview with respective staffs in HEP. The hard constraints for the UMSLIC HSAP are as follow:

1. (Hc_1) Student allocated in hostel buildings must be separated by gender.
2. (Hc_2) Student with physically challenged or disabled must be placed at first level of the hostel building.
3. (Hc_3) New intake students must be allocated in the campus hostel.
4. (Hc_4) The number of student allocated in each room must be less than or equal to the room capacity.

Meanwhile, the soft constraints for the UMSLIC HSAP are as follow:

1. (Sc_1) There must be student with different races in each unit of the hostel.
2. (Sc_2) Single room or privilege room must be allocated for those students who are highly active in institution curriculum activities and perform well in academic (i. e. cumulative grade point average (cgpa) of 3.5 and above).

In UMSLIC, the HEP is responsible for planning and managing the student's room allocation for each semester. Although HEP is producing a feasible solution for HSAP, the process of inspection student details and management of building facilities will take very long time to schedule and the results for the allocation process might not produce good quality, which may not fully satisfy the soft constraints Sc_1 or Sc_2. The quality of solution in HSAP can be calculated in terms of: (1) no violations of hard constraints, (2) satisfaction of additional requirements stated in soft constraints.

There are three types of hostel provided for the students: Alpha (female), Beta (male) and Mutiara (female). Each type of hostel consists of different number of blocks, floor, unit and room capacity. The allocation of the students to hostel is based on certain attributes such as gender, health status, demographic and participations in curriculum activities. Table 1 shows the number of rooms in each hostel.

Table 1. Summary of hostel attributes

Hostel	Alpha	Beta	Mutiara
Number of rooms	606	684	240

4 Mathematical Model of Problem Instance

In this section, the formal model of UMSLIC hostel allocation is presented. Some of the notations are listed as follows:

- Hostel, *H*: A group of building located at various location to be accommodate for students
- Block, *B*: Total number of buildings can be accommodate for students
- Floor, *F*: Level in each *B*
- Unit, *U*: Room unit in each *F*
- Room, *R*: Total number of rooms in *H*
- Students, *S*: Total number of students

All the blocks can be categorized as follow: $B = B_A + B_B + B_M$ where B_A is allocated for female student inside campus, B_B is allocated for male student inside campus and B_M is allocated for female student outside campus. For each type block, there are different numbers of blocks which can be formulated as follow:
$B_A = B_1 + B_2 + \ldots + B_n$, *where n is the total number of block in block type* B_A
$B_B = B_1 + B_2 + \ldots + B_n$, *where n is the total number of block in block type* B_B
$B_C = B_1 + B_2 + \ldots + B_n$, *where n is the total number of block in block type* B_M
whereby the total number of rooms are calculated based on each block.

In this domain, the objective is to minimize the cost function F_c as much as possible subject to hard constraints, $Hc = Hc_1 + Hc_2 + Hc_3 + Hc_4$ must be zero and soft constraints, $Sc = Sc_1 + Sc_2$ needs to be reduced. The cost function can be calculated as follow:

$$Fc = Hc + Sc \tag{1}$$

The penalty for hard constraint is set with the weight of 100,000 so that the cost function can easily identify if hard constraints exist in case the value is too high, noted that there are 4 hard constraints (λ_1 $until$ λ_4) while λ_5 and λ_6 are both soft constraints identified in UMSLIC. Meanwhile, the penalty for soft constraints is given based on the requirement stated by the HEP during interview session. λ_5 is given heavier penalty due to multi-racial is The penalty for every weight is summarized in Table 2.

Table 2. Penalty assignation of constraint violations.

Weight	Penalty	Description
λ_1 to λ_4	100,000	Hard constraints, *Hc*
λ_5	5	Soft constraint, Sc_1
λ_6	2	Soft constraints, Sc_2

The formulation model for both hard constraints and soft constraints are presented below. For hard constraint 1 (Hc_1) shown in equation (2), the violation when male and female students are not separated is given by

$$H_{C_1} = \lambda_1 \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} Ge_{ij} \tag{2}$$

where s is the total number of student, λ_1 is the weight, GE_{ij} is the violation if student i and student j is different gender.

For hard constraint 2 (Hc_2) shown in equation (3), the penalty given if the physically disabled or challenged student is not placed on the first floor is given by

$$H_{C_2} = \lambda_2 \sum_{i=1}^{f} \sum_{j=1}^{s} He_{ij} \tag{3}$$

where f is the total floor in a block, s is the total number of students, λ_2 is the weight, He_{ij} is the violation if physically challenged student j is placed at the floor i.

For the hard constraint 3 (Hc_3) in equation (4), the condition of first year student is not placed in the campus is given by

$$H_{C_3} = \lambda_3 \sum_{i=1}^{r} \sum_{j=1}^{s} Fr_{ij} \tag{4}$$

where r is the total number of rooms, s is the total number of students, λ_3 is the weight, FR_{ij} is violation if student j is the first year student which placed in room i which is located at outside the campus.

For the hard constraint 4 (Hc_4) shown in equation (5), the condition of the student allocated into the room which is larger than the capacity of that room is given by

$$H_{C_4} = \lambda_4 \sum_{i=1}^{r} \sum_{j=1}^{s} Ca_{ij} \tag{5}$$

where r is the total number of rooms, s is the total number of students, λ_4 is the weight, Ca_{ij} is the violation if student j assigned into room i which is fully occupied.

Meanwhile, in terms of soft constraint 1 (Sc_1) shown in equation (6), a given unit without multi-racial students is given by

$$S_{C_1} = \lambda_5 \sum_{i=1}^{u} \sum_{j=1}^{s} Rc_{ij} \tag{6}$$

where u is the total number of rooms, s is the total number of students, λ_5 is the weight, Rc_{ij} is the violation if student j allocated in the room unit i which has more than one similar race.

For soft constraint 2 (Sc_2) shown in equation (7), the single room or privilege room is allocated for normal student is given by

$$S_{C_2} = \lambda_6 \sum_{i=1}^{r} \sum_{j=1}^{s} Pr_{ij} \tag{7}$$

where r is the total number of rooms, s is the total number of students, λ_6 is the weight, Pr_{ij} is the violation when student j who is inactive and poor in academic performance but allocated single or privilege room i..

5 Stochastic Algorithms

There are two phases need to be considered in this research. The first phase deals with producing an initial solution using Constraint Programming while the second phase improves the quality from initial solution with both Great Deluge and Simulated Annealing algorithms. A five dimensional solution representation is created to represent the data structure of the solution. The hard constraints are identified through the means of diagonal matrices in the form of binary whereby each of the constraints are assigned based on the model stated in equation 2 to equation 5. Using the diagonal matrices, Constraint Programming perform movements search for feasible solution within restricted search space.

In the second phase of improvement, Simulated Annealing with linear cooling rate is employed and compared with both linear and non-linear Great Deluge algorithm. Simulated Annealing is introduced by Kirkpatrick [11] which derived from the process of heating and cooling metal in manufacturing in order to produce solid crystal structure. The cooling pattern in Simulated Annealing is well studied by many researchers [8, 15] which allow the searches to improve further. In this research, the parameters are adjusted such by continuous testing which allows Simulated Annealing performs differently and producing better result. At each reduction temperature, a neighbourhood search is performed with two types of movements:

- (N_1) Simple Search: moving an entity (student) to another feasible and available resources (room)
- (N_2) Swap Search: exchanging of resources (rooms) between two entities (students)

The neighbourhood solution will be accepted if the cost function (quality) is better. In this context, the lower the cost function, the better the quality of solution. If the neighbourhood solution is worse than the current solution, it can be accepted with a certain probability by using Boltzmann distribution with the given formula:

$$e^{\Delta Fc/T} \tag{8}$$

where ΔFc is the change in the quality between neighbourhood and current solution and T is the temperature. Generally, T is initialized with high temperature and slowly drop as the search is progressing. An initial temperature of 500 degrees are set in order to provide the algorithms to explore more in different search spaces. The cooling schedule is dropped by 0.05 in every reduction temperature. The cooling process is repeated until the temperature reach the floor temperature or the elapsed time meets the time limit. The detail of proposed Simulated Annealing is shown in Table 3.

Meanwhile, the implementation of Great Deluge is nearly similar with Simulated Annealing and hill climbing. The algorithm is invented by Dueck [5] which derived from the analogy of climbing hill to search for higher area in order to avoid getting drown when the water level rises. The algorithm also accepts worse solution if the cost function is less than or equal to the water level.

Table 3. Pseudocode for Simulated Annealing

```
Algorithm 1. Simulated Annealing
Generate initial solution from CP $s := s_0$, set temperature $t := t_0$, cooling rate, $\alpha$
repeat
    $s' := NeighbourhoodSearch(s \rightarrow (N_1 \text{ or } N_2))$
    $\delta := Fc(s') - Fc(s)$
    if$((\delta \leq 0) \text{ or } (exponential(-\delta/t) < random[0,1]))$
        $s := s'$
    endif
    $t := coolingStrategy(t)$
until termination condition $stop := true$
```

The neighbourhood movements adopted in Great Deluge are same with the neighbourhood performed in Simulated Annealing whereby N_1 or N_2 will be performed in every decrement of water level. The initial water level is set based on the *Fc* of initial solution produced by Constraint Programming and then increased by one hundred. As the water level decreases (in the case of minimization), the algorithm will force the solution to find another solution which is lower than the water level. The linear decay rate is considered to be standard version of Great Deluge while the improvise version using non-linear decay rate is introduced by Obit, J. H. [16, 18] which shows to produced promising result in course timetabling problem with benchmark datasets. The detail of proposed Great Deluge is shown in Table 4 and the non-linear decay rate is illustrated in Table 5.

Table 4. Pseudocode for Great Deluge

```
Algorithm 2. Great Deluge
Generate initial solution from CP $s := s_0$, set water level $\beta := Fc(s_0)$, decay rate, $\gamma$
set best solution, $s_{best} := s$
repeat
    $s' := NeighbourhoodSearch(s \rightarrow (N_1 \text{ or } N_2))$
    $\delta := Fc(s') - Fc(s)$
    if$((\delta \leq 0) \text{ or } (Fc(s') \leq \beta))$
        $s := s'$
    endif
    $\beta := decayRate(\gamma)$
until termination condition $stop := true$
```

The algorithm 2 shown in Table 4 performs linear reduction in γ. The parameter for the linear reduction in γ is set to 0.1 in order to allow the water level decrease slowly when the search is performed. Meanwhile, the pattern of non-linear decay rate is further described as shown in Table 5. The ε in Algorithm 3 (line 6) is a constant value of one. If the range between the water level and neighbourhood cost function is less than one, the water level is increased by two or else, the water level increase is based on the distribution of probability between the minimum and maximum value whereby the minimum and maximum value are 1 and 4 respectively.

Table 5. Pseudocode for non-linear decay rate in Great Deluge

```
Algorithm 3. Non-linear decay rate Great Deluge
Set range, $r := \beta - Fc(s')$
decay rate, $\gamma$ in Algorithm 2 (line 8)
if(range<1)
    $\beta := \beta + 2$
else
    $\beta := \beta \times (exponential(-\delta(random[min, max]))) + \varepsilon$
endif
```

6 Experimental Results

Two types of experiments are conducted in this research with each experiment using different instances. The first experiment is conducted with a benchmark dataset while the second experiment is using the real-world datasets from the institution campus, session 2016/2017. Each experiment is conducted with 50 runs and the average of the result is recorded. The algorithm stop improves the search when either one of the termination condition meets. Those conditions are:

- when the Simulated Annealing temperature reach floor temperature which is zero while the water level in Great Deluge is zero.
- when the time taken by the algorithm to perform improvement reach 300 seconds

The average cost is the cost function calculated based on equation 1 in Section 4.

Meanwhile the maximum cost and minimum cost indicates the best solution and worst solution produced by the algorithms throughout 50 runs. The percentage of improvements as shown in Table 6 indicates the improvement from initial solution (from Constraint Programming).

Table 6. Preliminary result for benchmark and real-world instance session 2016/2017

Datasets	Benchmark			UMSLIC session 2016/2017		
Algorithm	SA	GD		SA	GD	
Reduction	Linear	Linear	Non-linear	Linear	Linear	Non-linear
Average Cost, *Fc*	450.5	774.8	883.5	1178.5	1523.3	1473.4
Max cost, *Fc*	512	857	914	1217	1578	1501
Min cost, *Fc*	405	669	848	1159	1432	1445
Improvement, %	73.2	54.8	47.8	54.1	41.5	42.3

In this experiment, it is clearly shown (Table 6) that in average Simulated Annealing produced better solution compare to both linear and non-linear Great Deluge. This indicates there is huge difference between both techniques whereby the difference in total cost is more than 300. In terms of improvement of quality from initial solution, Simulated Annealing manages to improve the solution more than 70 percent while both linear and non-linear manage to improve 54.8 and 47.8 percent respectively. This implies that the linear reduction in Simulated Annealing manage to perform improvement much better as the temperature slowly decreases. Meanwhile, the next experiment is conducted towards the effectiveness of proposed algorithms in real-world problem using one of the dataset in UMSLIC institution. The overall results produced in real-world instance are much poorer compare with the benchmark problem. This indicates that much tighter constraints are involved in the real-world instance. Simulated Annealing with linear cooling reduction produced much better solution in average throughout 50 runs experiment compare to Great Deluge algorithm. Furthermore, the improvement done by Simulated Annealing is much higher compare to Great Deluge whereby Simulated Annealing manage to improve the solution more than 50 percent while Great Deluge only can improve the solution between 41 to 43 percent. This can be summarize that Simulated Annealing performs well in this problem domain.

7 Conclusion

This research has presented a stochastic approach into combinatorial optimization problem domain. Although there are very few studies on HSAP, this domain is considered an NP-hard optimization which is very similar with other SAP such as office allocation and timetabling problems. Many optimization algorithms are proposed into SAP is able to produce good solutions in their respective domains. This research also presented the formal mathematical model representation constraint in HSAP which focused in higher institution. Two stochastic approaches with different reduction parameters are investigated and conducted in preliminary studies. The experimental results are shown in Section 6. It can be stated that both stochastic algorithms are producing good solutions in each experiment. In this particular domain, Simulated Annealing outperformed Great Deluge. However, this does not imply that the same algorithm can perform well in other allocation problems as stated in [20].

For future research, more dynamic cooling strategies in Simulated Annealing can be study such as thermodynamic cooling which may be potentially producing much better solution than the current findings. Genetic Algorithm with different crossover operators can also be investigate in this domain as the literature findings show that Genetic Algorithm can perform well in HSAP [1].

References

1. Adewumi, A. O., & Ali, M. M.: A multi-level genetic algorithm for a multi-stage space allocation problem. Mathematical and Computer Modelling, 51(1-2), 109-126 (2010).
2. Bai, R.: An investigation of novel approaches for optimising retail shelf space allocation. Doctoral dissertation, University of Nottingham (2005).

3. Benjamin, C. O., Ehie, I. C., & Omurtag, Y.: Planning facilities at the University of Missouri-Rolla. Interfaces, 22(4), 95-105 (1992).
4. Burke, E. K., Cowling, P., Silva, J. L., & McCollum, B.: Three methods to automate the space allocation process in UK universities. In International Conference on the Practice and Theory of Automated Timetabling , pp. 254-273. Springer, Berlin, Heidelberg (2000).
5. Dueck, G.: New optimization heuristics: the great deluge algorithm and the record-to-record travel. Journal of Computational physics, 104(1), 86-92 (1993).
6. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers & operations research, 13(5), 533-549 (1986).
7. Junn, K. Y., Obit, J. H., & Alfred, R.: A Constraint Programming Approach to Solving University Course Timetabling Problem (UCTP). Advanced Science Letters, 23(11), 11023-11026 (2017).
8. Junn, K. Y., Obit, J. H., & Alfred, R.: Comparison of Simulated Annealing and Great Deluge Algorithms for University Course Timetabling Problems (UCTP). Advanced Science Letters, 23(11), 11413-11417 (2017).
9. Junn, K. Y., Obit, J. H., & Alfred, R.: The Study of Genetic Algorithm Approach to Solving University Course Timetabling Problem. In International Conference on Computational Science and Technology, pp. 454-463. Springer, Singapore (2017).
10. Kahar, M. N. M., & Kendall, G.: The examination timetabling problem at Universiti Malaysia Pahang: Comparison of a constructive heuristic with an existing software solution. European Journal of Operational Research, 207(2), 557-565 (2010).
11. Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P.: Optimization by simulated annealing. science, 220(4598), 671-680 (1983).
12. Landa Silva, J. D.: Metaheuristic and multiobjective approaches for space allocation. Doctoral dissertation, University of Nottingham (2003).
13. Landa-Silva, D., & Obit, J. H.: Comparing Hybrid Constructive Heuristics for University Course Timetabling (2011).
14. Obit, J. H., Alfred, R., & Abdalla, M. H.: A PSO Inspired Asynchronous Cooperative Distributed Hyper-Heuristic for Course Timetabling Problems. Advanced Science Letters, 23(11), 11016-11022 (2017).
15. Obit, J. H., Junn, K. Y., & Alfred, R.: A Performance Comparison of Metaheuristics Search for University Course Timetabling Problems. Advanced Science Letters, 23(11), 11012-11015 (2017).
16. Obit, J. H., Junn, K. Y., & Alfred, R.: Performance Comparison of Linear and Non-Linear Great Deluge Algorithms in Solving University Course Timetabling Problems. Advanced Science Letters, 23(11), 11027-11030 (2017).
17. Obit, J. H., Landa-Silva, D., Sevaux, M., & Ouelhadj, D.: Non-linear great deluge with reinforcement learning for university course timetabling. Metaheuristics–Intelligent Decision Making, Series Operations Research/Computer Science Interfaces, 1-19. Springer (2011).
18. Obit, J. H., Landa-Silva, D., Ouelhadj, D. & Sevaux, M.: Non-linear great deluge with learning mechanism for solving the course timetabling problem. In: 8th Metaheuristics International Conference. MIC (2009).
19. Pisinger, D.: An exact algorithm for large multiple knapsack problems. European Journal of Operational Research, 114(3), 528-541 (1999).
20. Wolpert, D. H., & Macready, W. G.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82 (1997).

Sensor Selection based on Minimum Redundancy Maximum Relevance for Activity Recognition in Smart Homes

Saed Sa'deh Juboor, Sook-Ling Chua, Lee Kien Foo

Faculty of Computing and Informatics, Multimedia University,
Persiaran Multimedia, Cyberjaya 63100, Malaysia
{sjb5609}@yahoo.com;{slchua,lkfoo}@mmu.edu.my

Abstract. Activity recognition in smart homes has attracted increasing attention from researchers due to its potential to recognize the occupant's activities of daily living such as showering, putting away laundry, grooming, etc. Recognizing the activities of daily living can help to support and assist the older adults, and enable them to continue living independently within their own homes. In order to support the occupants, activity recognition algorithms need to learn from a series of observations obtained from sensors. The central question that this paper aims to address is which sensors are informative for activity recognition. In this paper, the sensor selection problem is addressed using minimum-Redundancy Maximum-Relevance (mRMR) method.

Keywords: Activity Recognition, Sensor Selection, minimum-Redundancy-Maximum-Relevance, Information Gain, Mutual Information.

1. Introduction

Recognizing human daily activities within the smart homes is one of the important areas of research. A vital application for activity recognition is to support the older adults, who are living alone in the home. In order to learn about the occupant's daily activities, unobtrusive sensors such as state-change sensors are usually used. These sensors are attached to objects in the home such as doors, windows, cabinets, stoves, etc. A series of sensor observations is generated when the occupant carries out their daily activities such as opening cupboards and closing refrigerator door. An activity recognition system learns from the sensor observations to recognize the occupant's daily activities.

Among the challenges facing activity recognition system include choosing the right sensors to effectively recognize the occupant's activities. Many works reported in the literature, used a large number of sensors. However, some of these sensors could be redundant and/or irrelevant. A redundant sensor is a sensor that co-exists with another sensor, both being relevant to the activity recognition, but the removal of one of the sensor will not affect the recognition performance. For an example, in the case of sensors attached to the dryer and washing machine. The removal of either one of the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_24

sensor will not affect the recognition of "doing laundry" activity. An irrelevant sensor is a sensor that does not contribute towards distinguishing the different classes of activities. An example of this is the kitchen light, which could be involved in many activities such as preparing meals, washing dishes and making coffee. If other sensors can be used to distinguish between these activities, then the kitchen light sensor can be removed without affecting the recognition performance. When an activity recognition algorithm is trained on irrelevant and/or redundant sensors, it may affect the recognition performance and also need a larger training set.

The sensor selection problem is addressed using the minimum-Redundancy-Maximum-Relevance (mRMR) method. The amount of information obtained from each sensor is calculated and then a subset of sensors having the most correlation with an activity class and the least correlation between the sensors is selected. The effectiveness of our method is validated on two public datasets.

2. Related Work

Many methods have been proposed for sensor selection in the literature. There are two main approaches for sensor selection in smart homes: (i) wrapper-based and (ii) filter-based. The former involves a learning algorithm to assess the prediction performance of the sensors [1, 2]. However, the wrapper-based approach is time-consuming and computationally expensive, since the learning algorithm needs to search through the spaces of all sensor subsets and to be trained for each subset [1]. The filter-based method selects the intrinsic properties of the sensors, independent of any learning algorithms used. It results in a lower computational cost compared to the wrapper approach. Cook and Holder [3] employed the mutual information measure to rank and select important motion sensors for activity recognition. The mutual information measure is also applied in Dobrucalı and Barshan [4] to identify important sensors of wearable motion sensors for recognizing daily sport activities.

Chua and Foo [5] used information gain measure to select informative state-change sensors for recognizing activities of daily living. Rahman et al. [6] used information gain to identify eating periods depending on head motion captured from accelerometer, gyroscope, magnetometer and light sensors. Fahad et al. [7] applied information gain measure to select informative sensors on contact switch and motion sensors to improve the recognition performance of daily activities in a smart home. One of the limitations of information gain and mutual information methods is that they do not take into account of redundancy between sensors.

Other works combine wrapper-based and filter-based approaches for sensor selection such as the work of Fish et al. [8]. They used a wrapper-based decision tree algorithm and applied the mutual information at each node of a decision tree. Their work is applied on data obtained from accelerometers to recognize human physical activities. In similarity to our work, Chen et al. [9] used mRMR measure to identify useful sensors for energy usage prediction. However, their work is applied on motion sensors.

3. Proposed Method

There are many sensor selection methods that can be used to select relevant sensors for activity recognition. However, these methods do not consider the redundancy among sensors. The minimum-Redundancy Maximum-Relevance (mRMR) is a filter method that searches for relevance, i.e., dependence between the sensors and the activities, while also searching for redundancy, i.e., dependence among sensors.

Since our aim is to identify the set of sensors that have the most relevance (i.e., sensors that are highly correlated with an activity class) and minimal redundancy (i.e., sensors that have least correlation among other sensors), we can use mRMR for sensor selection. There are two conditions in mRMR, which calculates the relevance and redundancy of the sensors [10]:

i. Minimum-redundancy. This condition calculates the mutual information between sensors such that sensors selected are mutually dissimilar. The minimum-redundancy condition is thus:

$$min \quad R(X), R = \frac{1}{|X|^2} \sum_{S_i, S_j \in X} I(S_i, S_j) \tag{1}$$

where $I(S_i, S_j)$ represents the measure of mutual information between sensors S_i and S_j, X is the set of sensors and $|X|$ is the number of sensors in X .

ii. Maximal-relevance. This condition maximizes the relevance of the sensors S with respect to the activity class c:

$$max \quad D(S,c), D = \frac{1}{|X|} \sum_{S_i \in X} I(S_i, c) \tag{2}$$

where $I(S_i, c)$ is the mutual information between sensors S_i and activity class c.

By optimizing the above two conditions (Equations 1 and 2) simultaneously, we can therefore obtain the set of sensors having the most correlation (relevance) with the activity class and least correlation among the sensors (redundancy). The criterion function to optimize the two conditions is thus:

$$max = (D - R) \tag{3}$$

4. Dataset and Evaluation Method

The MIT PlaceLab dataset [11] and van Kasteren dataset [12] were used in this study. For the MIT PlaceLab dataset, 77 state-change sensors was used to collect information about the occupant. The occupant lived in the home for 16 days and recorded his activities. Since this paper aims to identify activities of daily living, there are five activities of daily living observed in this dataset i.e., toileting/showering, doing laundry, dressing/grooming, putting away/washing dishes, and preparing meals. Only 49 sensors were used to collect data of these five activities. A leave-two days-out cross validation method is used, where 14 days are used for training and 2 days for testing. This results in 8 training-testing splits. For the van Kasteren dataset [12], a total of 14 state-change sensors were installed in a three-room apartment. The occupant lived in the apartment over a period of 24 days. There were four activities observed in this dataset, which are leaving house, toileting/showering, going to bed and preparing meals. Since the number of activities and sensors used in this dataset were smaller compared to the ones in MIT PlaceLab, a leave-four days-out cross validation method is used, where 20 days are used for training and 4 days for testing. This results in 6 training-testing splits. The recognition accuracy is the main metric used in this study. The final recognition accuracy is computed by averaging the accuracies in each evaluation.

5. Experimental Results and Discussion

Two experiments were conducted on the benchmark datasets (discussed in Section 4). The proposed mRMR method is used for sensor selection in the first experiment. The second experiment compared the mRMR method with three baseline methods: (1) information gain, (2) mutual information and (3) full set of baseline sensors.

For each training set, 3 different classifiers were trained on the set of sensors selected. These classifiers are hidden Markov models (HMMs) [13], naïve Bayes classifier and *k*-nearest neighbor. The performance of the classifiers to recognize the occupant's activities is tested using the testing sets. The reason for using three different classifiers is to validate the results of sensor selection and not merely to compare among the classifiers and determine which performs better.

5.1 Experiment 1: Selecting sensors using mRMR method

The aim of this experiment is to test the effectiveness of mRMR for sensor selection. For each training set, the mRMR method is used for sensor selection. These sensors are then ranked based on their respective mRMR scores. A cut-off point is determined in order to select the set of informative sensors and eliminate sensors that are less important. If a higher cut-off point is chosen, the number of selected informative sensors will be insufficient for recognizing human activity. In view of this, two cut-off points were chosen for each dataset used. These cut-off points were chosen based on prior knowledge about the activities in the homes, so that the selected sensors will be representative for all activities.

For the MIT PlaceLab dataset, the two cut-off points are 0.065 and 0.1. Sensors with value less than the cut-off points are eliminated from both the train and test data sets respectively. It is important to eliminate these sensors which is as though they are removed physically from the home. This allowed us to test how well the set of informative sensors identified to recognize the occupant's activities. As a result, a total of 25 informative sensors was identified based on 0.065 cut-off point and a total of 14 informative sensors was identified based on 0.1 cut-off point.

As for the van Kasteren dataset, the two cut-off points are 0.08 and 0.165, resulting a total of 10 and 8 informative sensors identified. Since the number of sensors installed in the home and the number of activities to be recognized are different for both the datasets, the set of sensors selected is different as shown in Table 1 (MIT PlaceLab) and Table 2 (van Kasteren). The three classifiers are then trained on the selected set of sensors. The recognition results are shown in Table 3.

Table 1: Sensors selected using mRMR on MIT PlaceLab dataset

(a) Total of 25 sensors selected based on 0.065 cut-off point

Sensors identified for each activity				
Activity1 (Preparing meal)	**Activity 2 ((Grooming/ Dressing)**	**Activity 3 (Toileting/ Showering)**	**Activity 4 (Washing dishes)**	**Activity 5 (Doing laundry)**
S4 (Kitchen door)	S6 (Medicine cabinet)	S6 (Medicine cabinet)	S5 (Kitchen cabinet)	S29 (Laundry dryer)
S5 (Kitchen cabinet)	S7 (Medicine cabinet)	S7 (Medicine cabinet)	S16 (Dishwasher)	S47 (Kitchen Door)
S18 (Kitchen cabinet)	S14 (Bathroom cab net)	S14 (Bathroom cabinet)	S18 (Kitchen cabinet)	S48 (Washing machine)
S23 (Kitchen cabinet)	S15 (Sink faucet –hot)	S15 (Sink faucet –hot)	S23 (Kitchen cabinet)	
S27 (Kitchen Drawer)	S20 (Bedroom drawer)	S28 (Sink faucet cold)	S27 (Kitchen Drawer)	
S30 (Refrigerator)	S24 (Foyer closet)	S32 (Shower faucet)		
S42 (Kitchen Toaster)	S25 (Office drawer)	S37 (Toilet Flush)		
S44 (Freezer)	S28 (Sink faucet cold)	S41 (Bathroom door)		
	S31 (Office light switch)			
	S41 (Bathroom door)			
	S46 (Jewelry box)			

(b) Total of 14 sensors selected based on 0.1 cut-off point.

Sensors identified for each activity				
Activity1 (Preparing meal)	**Activity 2 ((Grooming/ Dressing)**	**Activity 3 (Toileting/ Showering)**	**Activity 4 (Washing dishes)**	**Activity 5 (Doing laundry)**
S4 (Kitchen door)	S6 (Medicine cabinet)	S6 (Medicine cabinet)	S5 (Kitchen cabinet)	S29 (Laundry dryer)
S5 (Kitchen cabinet)	S7 (Medicine cabinet)	S7 (Medicine cabinet)	S16 (Dishwasher)	S47 (Kitchen Door)
S18 (Kitchen cabinet)	S14 (Bathroom cab net)	S14 (Bathroom cabinet)	S18 (Kitchen cabinet)	S48 (Washing machine)
S23 (Kitchen cabinet)	S15 (Sink faucet –hot)	S15 (Sink faucet –hot)	S23 (Kitchen cabinet)	
S27 (Kitchen Drawer)	S20 (Bedroom drawer)	S28 (Sink faucet cold)	S27 (Kitchen Drawer)	
S30 (Refrigerator)	S24 (Foyer closet)	S32 (Shower faucet)		
S42 (Kitchen Toaster)	S25 (Office drawer)	S37 (Toilet Flush)		
S44 (Freezer)	S28 (Sink faucet cold)	S41 (Bathroom door)		
	S31 (Office light switch)			
	S41 (Bathroom door)			
	S46 (Jewelry box)			

Table 2: Sensors selected using mRMR on van Kasteren dataset

(a) Total of 10 sensors selected on 0.08 cut-off point

Sensors identified for each activity			
Activity 1 (Preparing meal)	**Activity 2 (Leave house)**	**Activity 3 (Toileting/ Bathing)**	**Activity 4 (Go to bed)**
S4 (Cups cupboard)	S7 (Front door)	S2 (Toilet door)	S14 (Bedroom door)
S5 (Fridge)		S3 (Bathroom door)	
S6 (Plates cupboard)		S9 (Toilet flush)	
S10 (Freezer)		S14 (Bedroom door)	
S13 (Groceries Cupboard)			

(b) Total of 8 sensors selected based on 0.165 cut-off point

Sensors identified for each activity			
Activity 1 (Preparing meal)	**Activity 2 (Leave house)**	**Activity 3 (Toileting/ Bathing)**	**Activity 4 (Go to bed)**
S5 (Fridge)	S7 (Front door)	S2 (Toilet door)	S14 (Bedroom door)
S6 (Plates cupboard)		S3 (Bathroom door)	
S13 (Groceries Cupboard)		S9 (Toilet flush)	
		S14 (Bedroom door)	

Table 3: Recognition accuracy results based on mRMR

(a) MIT PlaceLab Dataset

Test Set	Accuracy (%)					
	25 Sensors			14 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	77.32	84.54	89.69	84.62	85.9	85.90
2nd	80.77	82.35	82.69	82.50	82.5	82.50
3rd	71.43	78.18	82.14	71.79	73.68	73.68
4th	78.33	85.0	86.67	87.50	89.58	89.58
5th	80.82	93.15	90.41	91.07	92.85	92.86
6th	90.63	87.5	87.50	86.96	86.96	86.96
7th	88.30	91.49	89.36	90.14	90.0	90.14
8th	78.57	82.93	73.81	81.82	84.85	68.78
Avg	80.77	85.64	85.20	84.55	85.79	83.80

(b) van Kasteren Dataset

Test Set	Accuracy (%)					
	10 Sensors			8 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	100.00	79.49	97.44	100.00	79.49	94.87
2nd	100.00	90.23	100.00	100.00	89.66	100.00
3rd	100.00	88.64	100.00	100.00	88.64	100.00
4th	100.00	85.29	100.00	100.00	85.29	100.00
5th	100.00	88.00	100.00	100.00	88.00	100.00
6th	97.44	86.27	87.76	97.5	86.27	86.27
Avg	99.57	86.32	97.53	99.58	86.23	96.86

5.2 Experiment 2: Comparing mRMR Method with Baseline Methods

This experiment compared the mRMR approach with three baseline methods: (1) information gain, (2) mutual information and (3) full set of baseline sensors. The same process as seen in Experiment 1 is repeated in this experiment where we first applied the baseline methods for sensor selection. Sensors are then ranked according to their scores. In order to have a fair comparison between the baseline methods and our proposed method, we used two cut-off points in order to select sensors that have the equivalent number of sensors.

Information Gain Method. Information gain calculates the expected information for each sensor by selecting sensors that are highly correlated to activity class. The information gain, *Gain (A, S)* of sensor *S* relative to a collection of examples *A*, is:

$$Gain\ (A,S) = Entropy(A)\ -\sum_{v \in Values\,(S)} \frac{|A_v|}{|\,A\,|} Entropy\left(A_v\right) \tag{4}$$

where *Values(S)* is the set of all possible values for sensor *S* and A_v is the subset of *A*, for which sensor *S* has value v (i.e., $A_v = \{a \in A \,|\, S(a) = v\}$).

The first term in Equation 4 is the entropy of collection *A*, while the second term is the expected value of the entropy after *A* is partitioned using sensor *S*. The expected entropy described by second term is the sum of the entropies of each subset *A*, weighted by the fraction of examples that belong to *A*. *Gain (A, S)* is the expected reduction in entropy caused by knowing the value of sensor *S*. For the MIT PlaceLab datasets, we used 0.075 and 0.125 cut-off points, while for the van Kasteren dataset, we used 0.15 and 0.25 cut-off points. The results are shown in Table 4. Selecting sensors using mRMR (Table 3(a)) achieved an overall accuracy better than information gain (Table 4(a)) across all the three classifiers on MIT PlaceLab dataset. Both methods achieved the same recognition results on van Kasteren dataset.

Table 4: Recognition accuracy results based on information gain

(a) MIT PlaceLab Dataset

Test Set	Accuracy (%)					
	25 Sensors			14 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	78.50	82.80	88.71	82.19	84.93	87.67
2nd	76.47	82.35	78.00	80.56	77.78	77.78
3rd	72.00	76.00	80.00	71.43	74.29	74.29
4th	78.58	83.93	83.93	86.96	89.13	86.96
5th	84.85	92.65	89.71	87.50	91.07	89.29
6th	86.21	86.21	93.10	85.71	85.71	85.71
7th	85.19	91.36	91.25	90.63	90.63	89.06
8th	74.36	80.00	71.79	78.79	81.82	66.67
Avg	79.52	84.35	84.56	82.97	84.42	82.18

(b) van Kasteren Dataset

Test Set	Accuracy (%)					
	10 Sensors			8 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	100.00	79.49	97.44	100.00	79.49	94.87
2nd	100.00	90.23	100.00	100.00	89.66	100.00
3rd	100.00	88.64	100.00	100.00	88.64	100.00
4th	100.00	85.29	100.00	100.00	85.29	100.00
5th	100.00	88.00	100.00	100.00	88.0	100.00
6th	97.44	86.27	87.76	97.5	86.27	86.27
Avg	99.57	86.32	97.53	99.58	86.23	96.86

Mutual Information Method. Mutual information calculated the mutual dependence between the sensor and the target activity class. Therefore, sensors that have the largest dependency on the activity class will be selected for activity recognition. The mutual information between the sensor *S* and a set of activities *A* is:

$$MI(S, A) = \sum_{v \in Values(S)} \sum_{a \in (A)} log\left(\frac{P(v,a)}{P(v)\,P(a)}\right) \tag{5}$$

where *P(v,a)* is the joint probability distribution of *v* and *a*. *P(v)* and *P(a)* are the marginal distributions.

For MIT PlaceLab dataset, the cut-off points were 0.023 and 0.04, while for van Kasteren dataset, the cut-off points were 0.05 and 0.07. Sensors with value less than the cut-off points were eliminated from both datasets. The results are shown in Table 5. In comparison to the mutual information, when mRMR is used for sensor selection on the MIT PlaceLab dataset, the recognition results (Table 3(a)) achieved a slightly higher accuracy compared to mutual information (Table 5(a)).

As for van Kasteren dataset, sensor selection based on mRMR resulted in similar accuracy when using 10 informative sensors. When the classifiers were trained on 8 informative sensors, a higher accuracy (Table 3(b)) is achieved using mRMR compared to mutual information (Table 5(b)). A lower accuracy was observed with mutual information method is due to the reason being that sensor attached to the hall-bedroom door was not selected and thus “going to bed” activity was not recognized.

Table 4: Recognition accuracy results based on mutual information.

(a) MIT PlaceLab Dataset

Test Set	Recognition Accuracy (%)					
	25 Sensors			14 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	81.72	86.02	89.25	82.19	84.93	87.67
2nd	76.47	82.35	76.47	80.56	77.78	77.78
3rd	69.56	76.09	78.26	71.43	74.29	74.29
4th	78.58	82.14	83.93	86.96	89.13	86.96
5th	84.38	90.63	87.50	87.50	91.07	89.29
6th	82.76	86.21	93.10	85.71	85.71	85.71
7th	83.96	91.36	90.12	90.63	90.63	89.06
8th	75.0	80.00	72.50	78.79	81.82	66.67
Avg	79.05	84.35	83.89	82.97	84.42	82.18

(b) van Kasteren Dataset

Test Set	Recognition Accuracy (%)					
	10 Sensors			8 Sensors		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	100.00	79.49	97.44	64.10	64.10	64.10
2nd	100.00	90.23	100.00	82.76	82.76	82.76
3rd	100.00	88.64	100.00	81.82	81.82	81.82
4th	100.00	85.29	100.00	73.53	73.53	73.53
5th	100.00	88.0	100.00	84.00	84.00	84.00
6th	97.44	86.27	87.76	76.47	76.47	76.47
Avg	99.57	86.32	97.53	78.46	77.11	77.11

Full Set of Baseline Sensors. In this experiment, the three classifiers are trained directly on the full set of baseline sensors, without applying any sensor selection method. For the MIT PlaceLab dataset, the classifiers are trained on 49 sensors, while for the van Kasteren dataset, the classifiers are trained on 14 sensors. Table 6 presents the recognition results. The results showed that in the case of MIT PlaceLab dataset, when selecting sensors using the mRMR method (Table 3(a)), the average recognition accuracy using set of 25 sensors and set of 14 informative sensors is better than using the full set of 49 sensors across all the three classifiers. In the case of van Kasteren dataset, a similar accuracy is achieved when using 8 and 10 informative sensors. Figure 1 (MIT PlaceLab) and Figure 2 (van Kasteren) summarize the recognition results of the two experiments conducted. From the figures, our mRMR method clearly outperforms the baselines across all the classifiers.

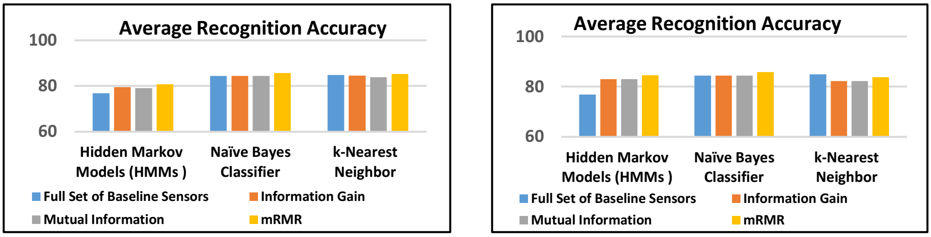
Table 4: Recognition accuracy results based on full a set of sensors.

(a) MIT PlaceLab Dataset

Test Set	Accuracy (%)		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	77.59	96.55	90.52
2nd	76.67	77.78	75.41
3rd	75.81	77.42	80.65
4th	74.63	81.82	83.33
5th	77.92	89.87	91.12
6th	85.29	81.82	87.88
7th	75.44	90.43	90.43
8th	71.43	79.59	79.59
Avg	76.85	84.41	84.87

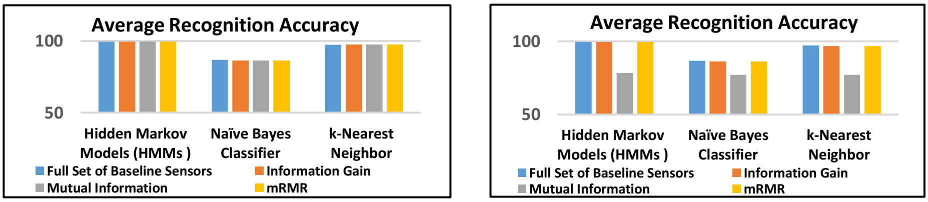
(b) van Kasteren Dataset

Test Set	Accuracy (%)		
	Hidden Markov Models (HMMs)	Naïve Bayes Classifier	*k*-Nearest Neighbor
1st	100	79.49	97.44
2nd	100	91.55	100
3rd	100	88.89	97.83
4th	100	85.29	100
5th	100	88.0	100
6th	97.44	87.04	88.24
Avg	99.57	86.71	97.25



(a) Set of 25 informative sensors (b) Set of 14 informative sensors

Fig. 1. Average recognition accuracy of the proposed method with baseline methods on MIT PlaceLab dataset: (a) Set of 25 informative sensors and (b) Set of 14 informative sensors.



(a) Set of 25 informative sensors (b) Set of 14 informative sensors

Fig. 2. Average recognition accuracy of the proposed method with baseline methods on van Kasteren dataset: (a) Set of 10 informative sensors and (b) Set of 8 informative sensors.

6. Conclusion

In this paper, the mRMR method is proposed for sensor selection. The effectiveness of the method is evaluated on two public datasets: MIT PlaceLab and van Kasteren. The proposed method is compared with three baseline methods: information gain, mutual information and full set of baseline sensors. The results showed that sensor selection based on mRMR method performed better than the baseline methods, and that it can effectively identify the important set of sensors for activity recognition.

Acknowledgments

This research is supported by the Ministry of Higher Education (MOHE), Malaysia under the Fundamental Research Grant Scheme (No: FRGS/1/2016/ICT02/MMU/02/3)

References

1. Koller, D., Sahami, M.: Toward Optimal Feature Selection. Stanford InfoLab, 284-292 (1996).
2. Chua, S.-L., Foo, L. K.: Tree alignment based on Needleman-Wunsch Algorithm for Sensor Selection in Smart Homes. Sensors, 17(8), 1902 (2017).
3. Cook, D., Holder, L.: Sensor selection to support practical use of health-monitoring smart environments. Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery, 1(4), 339-351 (2011).
4. Dobrucalı, O., Barshan, B.: Sensor-activity relevance in human activity recognition with wearable motion sensors and mutual information criterion. In Proceedings of the 28th International Symposium on Computer and Information Sciences, pp. 285–294. Paris, France (2013)
5. Chua, S. L., Foo, L. K.: Sensor Selection in Smart Homes. Procedia Computer Science, 69, 116-124 (2015).
6. Rahman, S. A., Merck, C., Huang, Y., Kleinberg, S.: Unintrusive eating recognition using Google Glass. In: 9th International Conference on Pervasive Computing Technologies for Healthcare, pp. 108-111. (2015).
7. Fahad, L.G., Tahir, S.F., Rajarajan, M.: Feature selection and data balancing for activity recognition in smart homes. In Proc. Of the IEEE Int. Conf. on Communications (ICC), pp. 512-517. London, UK (2015).
8. Fish, B., Khan, A., Chehade, N., Chien, C., Pottie, G.: Feature selection based on mutual information for human activity recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1729–1732. (2012).
9. Chen, C., Das, B., Cook, D.: Energy prediction in smart environments. In: Proceedings of the 6th International Conference on Intelligent Environments, pp. 148–154. IOS Press, (2010).
10. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell, 27(8), 1226–1238 (2005).
11. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors, in: Proceedings of Pervasive, pp. 158–175. (2004).

12. van Kasteren, T., Noulas, A., Englebienne, G., Kr¨ose, B.: Accurate activity recognition in a home setting. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 1–9. ACM (2008).
13. Juboor, S. S., Chua, S.-L., Foo, L. K.: Simultaneous Activity Segmentation and Recognition in Smart Homes. Journal of Engineering and Applied sciences, 11(8), 1851-1854 (2016).

Improving Network Service Fault Prediction Performance with Multi-Instance Learning

Leonard Kok[1], Sook-Ling Chua[1], Chin-Kuan Ho[1], Lee Kien Foo[1] and
Mohd Rizal Bin Mohd Ramly[2]

[1] Faculty of Computing and Informatics, Multimedia University, Cyberjaya 63100, Malaysia
1122700374@student.mmu.edu.my
{slchua, ckho, lkfoo}@mmu.edu.my
[2] Telekom Malaysia, Menara TM, Jalan Pantai Baharu, Kuala Lumpur 50672, Malaysia
mohdrizal.ramly@tm.com.my

Abstract. The internet has undoubtedly become everywhere essential to majority of the people from business services to entertainment. Early prevention of network service faults can greatly improve customer satisfaction and experience. Proactive network service faults prediction can certainly help the internet service providers to reduce service workloads and costs. One of the assumptions often made by standard supervised learning is to treat each session log (generated from the network management system) as an individual instance, where each instance is assigned a class label. Although such assumption is appropriate in some domains, it may not be appropriate in network service fault prediction since a network service fault is represented by a collection of session logs. In this paper, we aim to improve the network service fault prediction by transforming the single-instance learning to a multi-instance learning problem. We evaluate our proposed method on a real-world network data and compared with the baseline single-instance learning method. The multi-instance learning approach achieves a higher AUROC performance to single-instance learning approach.

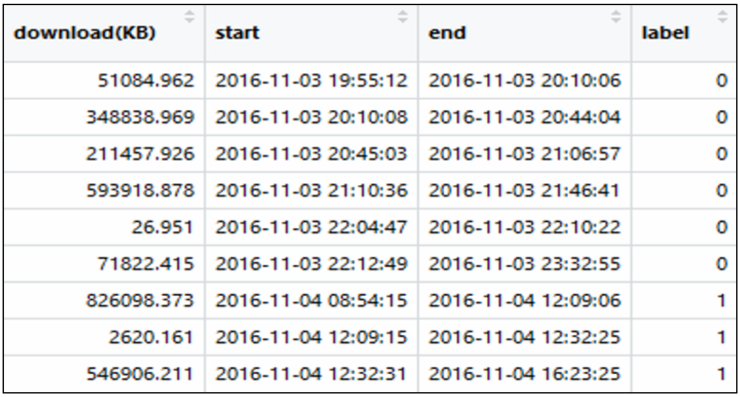
Keywords: Multi-instance learning, network service fault, multi-instance logistic regression.

1 Introduction

With the increasing number of internet subscribers, it has certainly put more strain on the internet service providers (ISP) to maintain a good quality of internet service. Ideally, keeping the internet connection available at all times would leave no complaints from the subscribers. However, network service faults and errors caused by equipment failures are bound to happen.

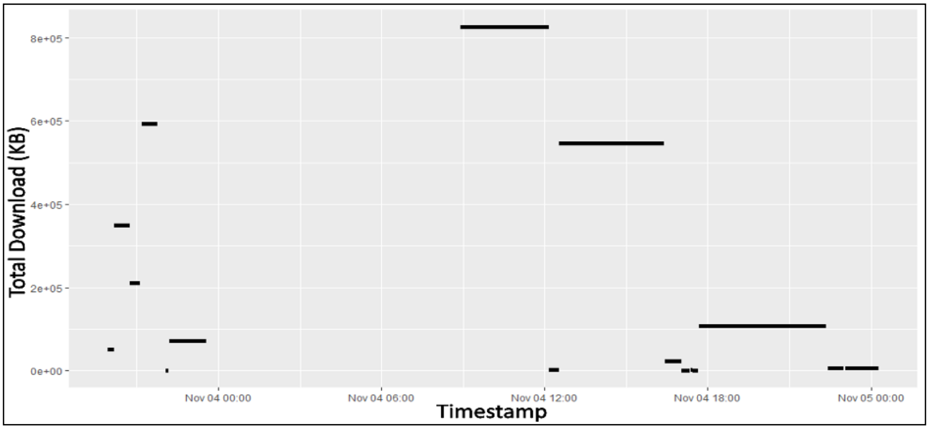
When these network service interruptions occurred, it is crucial for the ISP to react and resolve these issues as quickly as possible in order to maintain a good customer-support relationship. Yet reactive action alone from ISP is not enough to keep the internet service quality at a satisfactory level. Before any subscriber raises concern, ISP should proactively examine and repair network service faults to greatly reduce the internet service downtime [1, 2].

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_25



download(KB)	start	end	label
51084.962	2016-11-03 19:55:12	2016-11-03 20:10:06	0
348838.969	2016-11-03 20:10:08	2016-11-03 20:44:04	0
211457.926	2016-11-03 20:45:03	2016-11-03 21:06:57	0
593918.878	2016-11-03 21:10:36	2016-11-03 21:46:41	0
26.951	2016-11-03 22:04:47	2016-11-03 22:10:22	0
71822.415	2016-11-03 22:12:49	2016-11-03 23:32:55	0
826098.373	2016-11-04 08:54:15	2016-11-04 12:09:06	1
2620.161	2016-11-04 12:09:15	2016-11-04 12:32:25	1
546906.211	2016-11-04 12:32:31	2016-11-04 16:23:25	1

(a) Session logs data structure from one of the users. Each row represents a session log instance with class label.



(b) Session logs plot from one of the users. Each horizontal bar represents a session log instance interval. The beginning and ending of each horizontal bar represents session starts time and ends time.

Fig. 1. Network session logs with labels

Such proactive action can be taken through examination of the subscribers' network usage data (equipment generated session logs) such as total download bytes, total upload bytes, session start and end time, etc., and correlating the session logs with the customer ticket records to identify the faulty cases. A session log could be generated by:

- User requests, where the user turns off the premise's equipment. This would terminate the current session with a recorded end time.

- Self-termination, where the premise's equipment disconnects and reconnects to the network without any user intervention. This would terminate the current session and generate a new session instance within seconds.

As a daily log data could consist of hundreds of thousands of instances, it is rather time consuming for the domain experts to manually examine the network data. To reduce the experts' workload and improve service quality level, ISP are relying on supervised learning, which learns from historical service disruptions in order to predict potential anomalous sessions.

In supervised learning, the algorithms learn concepts from labeled examples. A set of example cases with each being labeled is fed as input to the algorithm, which will then predict labels of future examples. This means that each network session log instance is assigned a class label, either 1 (i.e., "faulty") or 0 (i.e., "non-faulty") as shown in Figure 1. However, learning under such assumption may not be appropriate since a network service fault could be represented by a collection of session logs that are related to one another temporally rather than fault as a single-instance level.

Keeping this in mind, we view the network data as a multi-instance (MI) learning, since some of the instances (session logs) are related as if they were from the same event. Our aim in this paper is to enhance the prediction performance of network service fault using multi-instance learning, where each example consists of a multiset (bag) of instances.

This paper is organized as follows: Section 2 discusses the related work. Section 3 describes the proposed methodology based on multi-instance learning and Section 4 discusses the experimental setup. Section 5 presents the results and discussion, and conclusion in Section 6.

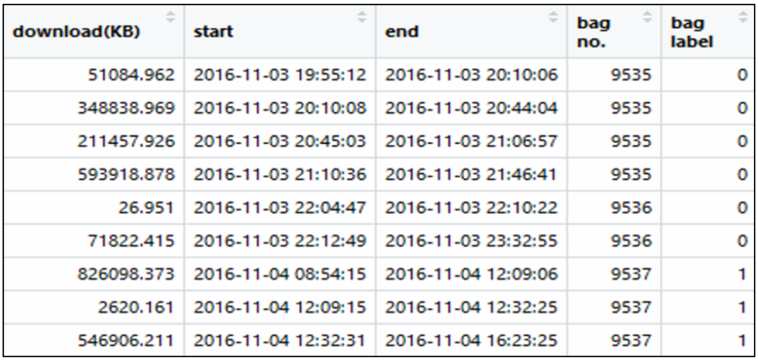
2 Related Work

There are many studies on multi-instance learning in computer vision (such as object recognition, image categorization, etc.) and bioinformatics (such as drug activity prediction, classifying patient phenotype, identification of bioactive conformers, etc.) domains. Dietterich et al. [3] proposed multi-instance (MI) learning to solve the ambiguous training examples with multiple alternative representations, such as the different shapes from a single molecule. Every combination of different bonds angles from a molecule can be grouped together as a training bag for MI learning.

Chen and Wang [4] used MI learning for region-based image categorization. In their work, the training and testing bags consist of compound objects where a bag label is determined by different class that satisfied various properties. An image could be labeled as 'mountains' and 'glaciers' as it contains both regions.

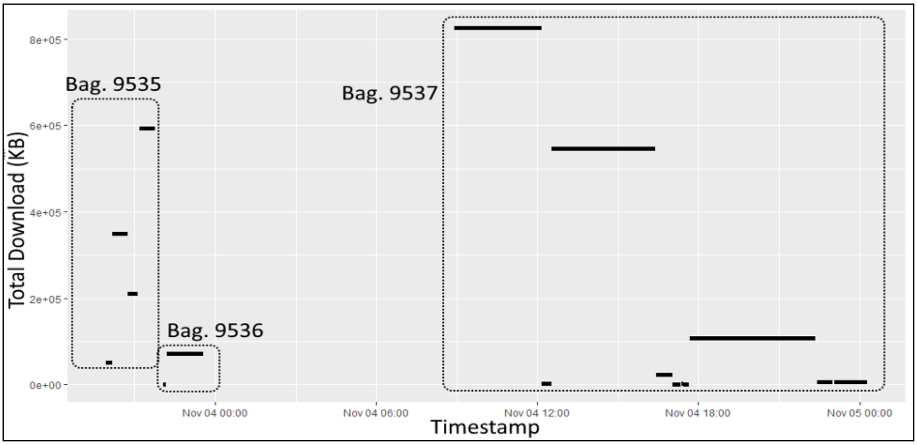
Popescu and Mahnot [5] proposed using MI learning to recognize illness on in-home monitoring sensors. They created bags that similar to evolving object where each bag consists 24 vectors of sensor data that corresponds to 24 hours window. While there is a similarity between their work and our work where the bag is labeled by corresponding to user's report, the condition that we generate bags from instances is different to their approach.

An empirical comparison between supervised learning (single-instance learning) and MI learning has been explored by Ray and Graven [6]. They showed that MI algorithms consistently outperform its single-instance learning counterparts on MI-nature data such as MUSK. In contrast, our work aims to improve the prediction performance by introducing bag-based labeling to our network data and create MI-based models based on the bags.



download(KB)	start	end	bag no.	bag label
51084.962	2016-11-03 19:55:12	2016-11-03 20:10:06	9535	0
348838.969	2016-11-03 20:10:08	2016-11-03 20:44:04	9535	0
211457.926	2016-11-03 20:45:03	2016-11-03 21:06:57	9535	0
593918.878	2016-11-03 21:10:36	2016-11-03 21:46:41	9535	0
26.951	2016-11-03 22:04:47	2016-11-03 22:10:22	9536	0
71822.415	2016-11-03 22:12:49	2016-11-03 23:32:55	9536	0
826098.373	2016-11-04 08:54:15	2016-11-04 12:09:06	9537	1
2620.161	2016-11-04 12:09:15	2016-11-04 12:32:25	9537	1
546906.211	2016-11-04 12:32:31	2016-11-04 16:23:25	9537	1

(a) Bagged session logs data structure from one of the users. "Bag no." indicates which bag the instance belongs to and "bag label" is the class label with 0 as non-faulty and 1 as faulty.

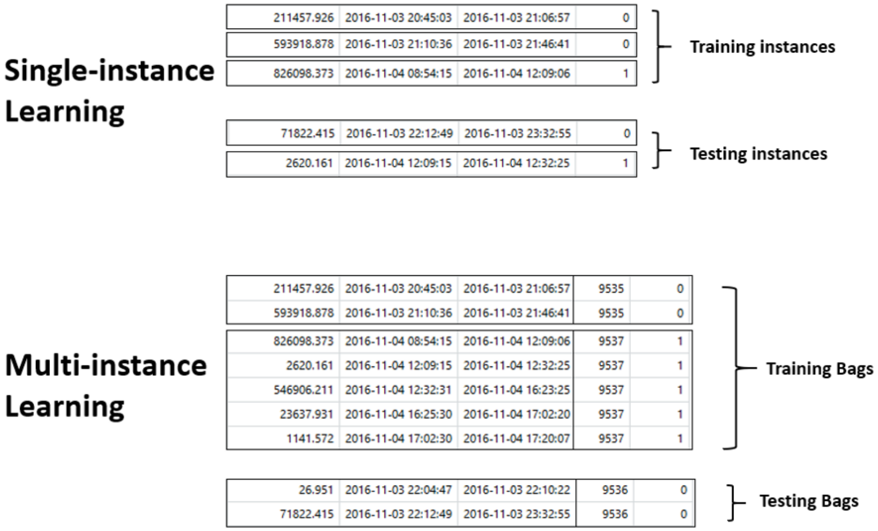


(b) Bagged session logs plot. Each dotted square box represents a bag. A total of 3 bags (Bag.9535, Bag.9536, Bag.9537), with each bag having different number of session log instances.

Fig. 2. Bagged network session logs with labels

3 Proposed Methodology

In this paper, we view the network service fault prediction as a multi-instance learning problem due to the reason being that a faulty event consists of several instances of session logs that are temporally related (rather than based on single-instance). In multi-instance learning, the algorithms learn concepts from labeled bags, consisting of multiset of instances. The goal is to predict labels of future bags. Figure 2 shows a sequence of session instances, which has been bagged with class label.



Single-instance Learning

Training instances

211457.926	2016-11-03 20:45:03	2016-11-03 21:06:57	0
593918.878	2016-11-03 21:10:36	2016-11-03 21:46:41	0
826098.373	2016-11-04 08:54:15	2016-11-04 12:09:06	1

Testing instances

71822.415	2016-11-03 22:12:49	2016-11-03 23:32:55	0
2620.161	2016-11-04 12:09:15	2016-11-04 12:32:25	1

Multi-instance Learning

Training Bags

211457.926	2016-11-03 20:45:03	2016-11-03 21:06:57	9535	0
593918.878	2016-11-03 21:10:36	2016-11-03 21:46:41	9535	0
826098.373	2016-11-04 08:54:15	2016-11-04 12:09:06	9537	1
2620.161	2016-11-04 12:09:15	2016-11-04 12:32:25	9537	1
546906.211	2016-11-04 12:32:31	2016-11-04 16:23:25	9537	1
23637.931	2016-11-04 16:25:30	2016-11-04 17:02:20	9537	1
1141.572	2016-11-04 17:02:30	2016-11-04 17:20:07	9537	1

Testing Bags

26.951	2016-11-03 22:04:47	2016-11-03 22:10:22	9536	0
71822.415	2016-11-03 22:12:49	2016-11-03 23:32:55	9536	0

Fig. 3. Example training and testing data between single-instance learning and multi-instance learning

The question, however, is how to construct the bag? The bags can be constructed through the combination of multiple instances based on the sequence and temporal closeness between session log instances. The constraint for the bagging process that we used to determine which bag should the instance belongs to is by comparing between the *i* instance's start time and *i*-1 instance's end time.

From our observations and with prior knowledge of the network data, a threshold of 5-minutes window is used to determine whether an instance belongs to the same bag. In other words, if the user turns off his/her premise's equipment for more than 5 minutes, then we would consider the next session logs entry as another new event (bag).

Figure 2 shows an example of generated bagged data based on 5-minute window. A label is assigned to each bag following the standard multi-instance assumption where a bag is labeled as positive (“faulty”) if there is at least one positive instance in the bag [7].

Figure 3 compares the single-instance learning and multi-instance learning. In single-instance (SI) learning, each learning object is described by an individual instance of session log, while in multi-instance learning, each learning object is described by a bag, consisting of multiset instances of session logs.

4 Experimental Setup

In this section, we describe the experimental setup of our work.

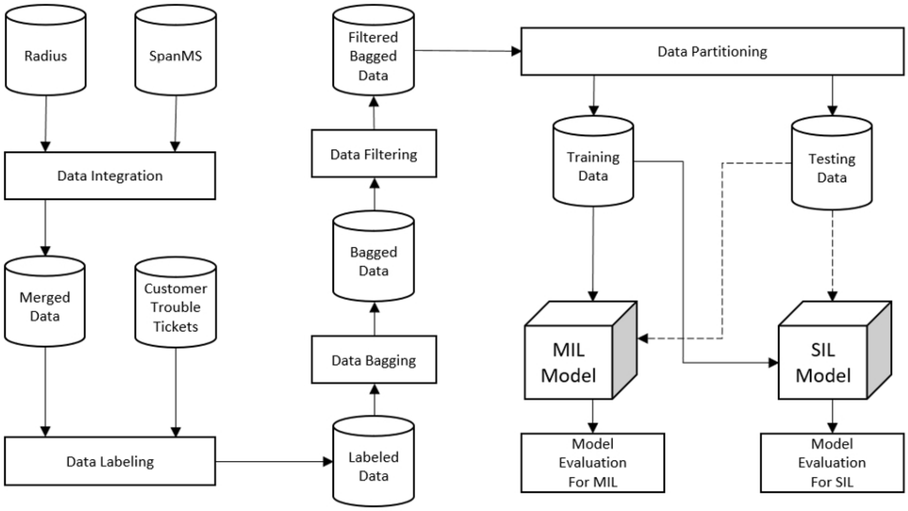


Fig. 4. Flow chart of the experimental setup

4.1 Data Integration and Data Labeling

As shown in figure 4, we obtained 3 types of network data (RADIUS, SpanMS, Customer Trouble Tickets (CTT)) from a telecommunication company in Malaysia. RADIUS data capture information about each subscriber’s behaviors of using the network such as the time they connect/disconnect to/from the network, the duration they spent on the network, download/upload bytes. SpanMS data provide insights of the hardware status such as the optical network units (ONU) and optical network terminals (ONT). Both RADIUS and SpanMS data are merged together via subscriber ID. In order to label the data, we need to correlate the merged data with CTT since CTT data provide information about when network service faults occurred. We followed the approach taken in [8] for data labeling.

4.2 Data Bagging

We bagged the instances by applying the proposed 5-minutes window. We labeled each bag with positive ("faulty") or negative ("non-faulty") class following the standard MI assumption, i.e., the bag is labeled as 'positive' as long as there is one positive instance in the bag.

4.3 Data Filtering

To ensure that our experiments have adequate bags for training and testing, and are not affected by the imbalanced class problem, we select users that have at least 50 bags and have a closely balanced class bags (such that the no. of negative ("non-faulty") bags is not greater than twice of the no. of positive ("faulty") bags). A total of 40 users was selected.

4.4 Data Partitioning

We partitioned the 40 selected users' data into 70% for training and 30% for testing. For the multi-instance learning, we used the bag-level labels for learning and thus all instance-level label was removed from the training and testing data respectively. As for the single-instance learning, we removed the bag-level labels and use the instance-level label for learning.

4.5 Model building

For this experiment, we used boosted logistic regression to build single-instance learning model and multi-instance logistic regression for multi-instance learning. The logistic regression is a commonly used probabilistic model for supervised learning of data with binary response variable [9]. Since our interest is to predict network service fault, the logistic regression is applicable and well-suited in this context. We proposed the use of boosted logistic regression, introduced by Friedman et al. [10], which is less susceptible to the overfitting problem [11]. The boosted logistic regression (BLR) method is used as a baseline of single-instance learning.

In order to perform a fair comparison, we used the multi-instance logistic regression (MILR), proposed by Xin and Frank [12] as the method for multi-instance learning.

4.6 Model Evaluation

The performance measurements that we used to evaluate the algorithms' performance are recall, precision, F1-score and AUROC.

Recall, also known as true positive rate, measures the number of positive instances or bags that a classifier predicted correctly. Precision, also known as positive predictive value, measures the number of predicted positive instances or bags correctly classified. While the F1-score calculates the harmonic mean between recall and precision. Furthermore, we also calculate the area under the ROC curve (AUROC), which tells us the tradeoff between recall and specificity (false positive rate).

Table 1. Average recognition performance of our proposed multi-instance logistic regression (MILR) and baseline single-instance boosted logistic regression (BLR)

Algorithm	Recall	Precision	F1-score	AUROC
MILR	0.67	0.57	0.60	0.71
BLR	0.58	0.54	0.53	0.65

Table 2. *p*-values of paired *t*-test

Recall	Precision	F1-score	AUROC
0.0141	0.2130	0.0382	0.0161

5 Results and Discussion

Table 1 shows the average F1-score, recall, precision and AUROC of 40 users. From the table, we can see that MILR multi-instance learning approach has a better recognition performance compared to BLR single-instance learning across all the 4 performance measurements. Specifically, the mean of recall rate increased from 0.58 to 0.67 using MILR. As for the precision, MILR achieved a slightly higher mean value compared to BLR. The improvement of recall rate and precision of MILR showed that multi-instance learning can effectively identify more relevant positive ("faulty") samples from the testing dataset, while maintaining the relevancy in the selected positive samples. From the results of F1-score and AUROC, it also supports that MILR method has a better performance in detecting network service fault. This is important to enable early fault detection and reduce maintenance service workloads.

We have also conducted a significant test and the *p*-values of paired t-test is shown in Table 2. The *p*-values show significant improvement in F1-score, recall and AUROC. This means that our method based on multi-instance learning effectively improves the network service fault prediction.

6 Conclusion

In this paper, we improved the prediction of network service disruptions based on multi-instance learning. We have also shown how individual session instances are bagged as input to the MIL algorithm. We have evaluated our method on real-world network data and compared with the baseline single-instance learning. The results showed that our method based on multi-instance learning effectively improves the prediction of network service disruptions. For the future work, we plan to fine-tune the time windows by intelligently determine the size of the window during the data bagging process in order to create a more accurate event or bag-based dataset for model training and testing.

Acknowledgments

This research is supported by the Ministry of Higher Education (MOHE), Malaysia under the Fundamental Research Grant Scheme (No: FRGS/1/2016/ICT02 /MMU/02/3).

References

1. Terzi, Duygu Sinanc, Ramazan Terzi, and Seref Sagiroglu. "Big data analytics for network anomaly detection from netflow data." In Computer Science and Engineering (UBMK), 2017 International Conference on, pp. 592-597., (2017).
2. Tran, An Trung. "Network Anomaly Detection." Future Internet (FI) and Innovative Internet Technologies and Mobile Communication (IITM) Focal Topic: Advanced Persistent Threats 55, (2017).
3. Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." Artificial intelligence 89, no. 1-2: 31-71., (1997).
4. Chen, Yixin, and James Z. Wang. "Image categorization by learning and reasoning with regions." Journal of Machine Learning Research 5: 913-939., (2004).
5. Popescu, M., and A. Mahnot. "Early illness recognition using in-home monitoring sensors and multiple instance learning." Methods of information in medicine 51, no. 4: 359., (2012).
6. Ray, Soumya, and Mark Craven. "Supervised versus multiple instance learning: An empirical comparison." In Proceedings of the 22nd international conference on Machine learning, pp. 697-704. ACM, (2005).
7. Herrera, Francisco, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans. "Multiple instance learning." In Multiple Instance Learning, pp. 17-33., (2016).
8. Leonard Kok, Sook-Ling Chua, Chin-Kuan Ho, Lee Kien Foo, and Mohd Rizal Bin Mohd Ramly. "Annotating Network Service Fault Based on Temporal Interval Relations." In International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, pp. 83-92., (2017).
9. Bewick, Viv, Liz Cheek, and Jonathan Ball. "Statistics review 14: Logistic regression." Critical care 9, no. 1: 112., (2005).
10. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics 28, no. 2: 337-407., (2000).
11. Mayr, Andreas, Harald Binder, Olaf Gefeller, and Matthias Schmid. "The evolution of boosting algorithms." Methods of information in medicine 53, no. 06: 419-427., (2014).
12. Xu, Xin, and Eibe Frank. "Logistic regression and boosting for labeled bags of instances." In Pacific-Asia conference on knowledge discovery and data mining, pp. 272-281., (2004).

Identification of Road Surface Conditions using IoT Sensors and Machine Learning

Jin Ren Ng[1], Jan Shao Wong[2], Vik Tor Goh[1](✉), Wen Jiun Yap[1],
Timothy Tzen Vun Yap[2], and Hu Ng[2]

[1] Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
[2] Faculty of Computing & Informatics, Multimedia University, Cyberjaya, Malaysia
{ngjinren, janshao.terry}@gmail.com,
{vtgoh, wjyap, timothy, nghu}@mmu.edu.my

Abstract. The objective of this research is to collect and analyse road surface conditions in Malaysia using Internet-of-Things (IoT) sensors, together with the development of a machine learning model that can identify these conditions. This allows for the facilitation of low cost data acquisition and informed decision making in helping local authorities with repair and resource allocation. The conditions considered in this study include smooth surfaces, uneven surfaces, potholes, speed bumps, and rumble strips. Statistical features such as minimum, maximum, standard deviation, median, average, skewness, and kurtosis are considered, both time and frequency domain forms. Selection of features is performed using Ranker, Greedy Algorithm and Particle Swarm Optimisation (PSO), followed by classification using k-Nearest Neighbour (k-NN), Random Forest (RF), and Support Vector Machine (SVM) with linear and polynomial kernels. The model is able to achieve an accuracy of 99%, underlining the effectiveness of the model to identify these conditions.

Keywords: Road surface conditions · Classification · Machine learning.

1 Introduction

Poorly maintained roads have been known to be causes of vehicle damage and road accidents. In most cases, roads containing potholes, cracks, or debris are just annoyances that are best avoided but in some severe cases, these hazards can result in fatal road accidents. For example, in November 2016, two cousins were killed when the car that they were travelling in hit a pothole, causing it to skid and then overturn [3].

The problem is not so much the lack of funds to maintain roads. In fact, governments have set aside vast amount of money just for this reason. In Malaysia, the federal government has set aside RM1 billion in its 2014 budget to maintain all state roads [2]. The actual challenge in road maintenance is the inability to monitor the incredibly vast road networks in a comprehensive and timely manner. For example, most councils carry out manual road inspections or only respond whenever complaints from the general public are received. Although

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_26

there are special road survey vehicles designed just for this purpose, they are expensive and difficult to use. Clearly, current approaches show a need for a simpler and more cost-effective method to monitor roads.

In our work, we propose to attach Internet-of-Things (IoT) sensors onto vehicles as data collection devices. Data collected from these sensors are aggregated and analysed using machine learning algorithms to automatically classify the various road surface conditions such as potholes, speed hump, or uneven roads. With the system, councils will now have a cost-effective way to monitor roads, which in turn allows for faster repairs and ultimately, improve the overall safety of roads.

2 Related Work

Current road surface condition monitoring systems utilise smartphones due to ubiquity, versatility, and enhanced computing power. Not only that, most smartphones have useful built-in sensors such as accelerometer and gyroscopes. There are numerous work which uses smartphones as data collection platforms such as those by Mohan et al. [11], Perttunen et al. [12], Bhoraskar et al. [6], Astarita et al. [5], and Alqudah [4]. Although they have been shown to work reasonably well, most of these work have assumed that smartphones are placed in a fixed orientation on the vehicle's dashboard. More often than not, smartphones are left in pockets or handbags of users. The effects of dampening due to the vehicle's suspensions, human body, and cushioned seats further reduce the sensitivity of the smartphone's sensors, thus making the detection of road surfaces less accurate. Besides that, smartphone-based solutions require that an app be left constantly on, which in turn affects the battery life. Battery life reduces even faster when other vital sensors such as accelerometer, gyroscope and GPS are used.

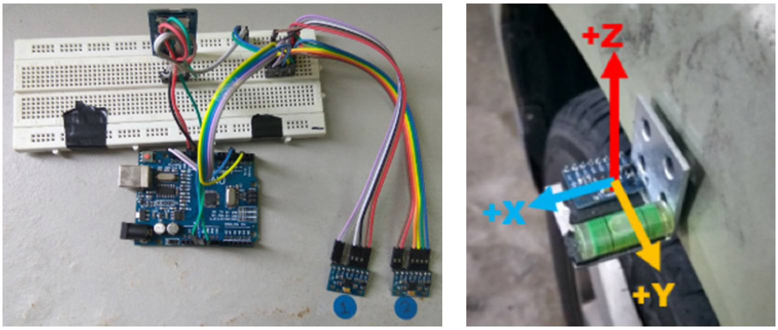
An alternative to smartphone-based solutions is to use specialised sensors as data collection devices. These are often small sensors which are semi-permanently attached to the vehicle. In fact, most of the earliest works use sensors such as these, including CarTel [9], Busnet [8], and more recently systems by Sun et al. [13], Collotta et al. [7] as well as Mednis et al. [10]. These systems extract and analyse data from the sensors and any high-energy events that are recorded by the accelerometer in a moving vehicle (e.g. sudden changes in acceleration in the Z-axis) are labelled as road anomalies, thus indicating the presence of either a bump or pothole. However, most of these work rely on traditional statistical or threshold-based approach in determining the road surface conditions. Each road condition requires a model be developed and this can be lengthy and tedious process. Furthermore, the model is neither dynamic nor adaptive. It cannot dynamically adapt to different types of roads (e.g. trunk road, highway) and their corresponding traffic volume.

To the best of our knowledge, there is no road surface monitoring system that takes advantage of the connectivity of IoT sensors while leveraging on versatility of modern machine learning algorithms that are easy to implement and highly adaptable to new scenarios. We will examine the how machine learning algorithms can further improve the detection and classification of various road surfaces.

3 Design and Implementation

3.1 Data Collection

For this project, the IoT sensors used to collect data are designed and implemented on the Arduino microcontroller platform. Specifically, we used the Arduino Uno together with two MPU6050 3-axis accelerometers. The hardware setup can be seen in Fig. 1a. The accelerometers are placed by the front wheels of a four-wheeled vehicle, both on the left and right side of the vehicle as shown in Fig. 1a. These positions were chosen because any contact made between the vehicle and the road will be felt most strongly near the wheels. Fig. 1b also shows the orientation of the accelerometers relative to the car. X-axis measures left and right motions, Y-axis measures forward and backward movements, and Z-axis measures vertical movements.



(a) Arduino with accelerometers (b) Sensor's orientation

Fig. 1. Hardware setup of road monitoring sensors.

Once the sensor have been installed on the vehicle, we proceeded to record the accelerometer readings for various road conditions. In particular, we measured the following 7 road surface conditions; (1) smooth road, (2) speed hump, (3) pothole, (4) rumble strip, (5) uneven road, (6) sudden stops, and (7) turns. To ensure sufficient variance for each road condition, we identified 5 different locations for each road condition. For instance, we measured 5 different stretches of smooth roads at locations A, B, C, D, and E. Finally, 10 readings are taken for each road condition at each location to ensure the consistency and sufficiency of data. If we deem that one *set* of data consists of the X, Y, and Z accelerometer readings of one road condition per location, we have a total of:

$$10 \text{ readings per condition} \times 5 \text{ locations} \times 7 \text{ road conditions} = 350 \text{ sets}$$

All readings are taken when the vehicle was driven at a speed of 20 km/hr. Results and discussion of these results are presented in Section 4.

3.2 Feature Extraction and Selection

The raw acceleration data from the MPU6050 accelerometers provided readings in the units of g-force or simply g. These readings are then processed to extract more useful statistical features. The extracted features are shown in Table 1.

Table 1. List of statistical features.

Domain	Statistical Parameters	Domain	Statistical Parameters
Time	Minimum	Frequency	Minimum
	Maximum		Maximum
	Standard Deviation		Standard Deviation
	Median		Median
	Mean		Mean
	Skewness		Skewness
	Kurtosis		Kurtosis
	Absolute Skewness		Absolute Skewness
	Absolute Kurtosis		Absolute Kurtosis
	Energy		Energy

Upon obtaining the statistical features from the raw data, we further applied various feature selection algorithms to optimise the number of features to be used in the classification algorithm. In this work, we tested three feature selection techniques, namely Ranker, Greedy Algorithm, and Particle Swarm Optimisation (PSO).

Briefly, the Ranker technique ranks features by their individual evaluations, hence its name. Higher ranked features are kept in the dataset while lower ranked ones are removed. On the other hand, the Greedy Algorithm is a greedy forward or backward search through the space of features. Finally, the PSO technique explores a feature's space using the PSO algorithm. This computational method optimises a problem by trying to improve a candidate solution given the measure of a quality and does so iteratively. The results of the feature selection algorithms are shown and discussed later in Section 4.2.

3.3 Classification

In order to classify the various road surface conditions into their respective categories, we tested three classifiers, namely (1) k-Nearest Neighbour (k-NN), (2) Random Forest (RF), and (3) Support Vector Machine (SVM). For k-NN, we varied the parameter k which represents the number of neighbours in order to empirically obtain the most optimum value. On the other hand, for RF, three types of parameters are tested to determine the most optimal results. These parameters are seeds (s), number of iterations (i), and number of features (f). Finally, we use both the Linear and Polynomial kernels for SVM. The Linear kernel has a parameter called cost (c) which can be varied to obtain different results. The Polynomial kernel has four parameters which are cost (c), coefficient (r), degree (d), and gamma (γ).

We used the machine learning software called Waikato Environment for Knowledge Analysis or more commonly known as Weka [1]. Weka is written in Java and contains a collection of visualisation tools and algorithms for data analysis and predictive analysis, hence simplifying and facilitating the development of new machine learning schemes.

We used Weka to carry out and evaluate the effectiveness of these classifiers in categorising the various road conditions. Before we can analyse the data, we applied the normalisation filter first, thus normalising all numeric values to a range of [0,1]. Subsequently, we applied the three classifiers on the normalised data which gave us an analysis of their respective performances. The performances of these classifiers are shown in Section 4.2.

4 Results and Discussions

4.1 Accelerometer Readings for Various Road Surface Conditions

The accelerometer readings for the various road conditions are shown in Fig. 2. Due to space constraints, of the 7 road surface conditions that we listed earlier in Section 3, we only present 5 road conditions and an additional control reading of a idling car (engine is running but vehicle is stationary). In addition, we have averaged out the numerous readings for each road condition so that they can be concisely presented in this paper.

The control reading of an idling vehicle sets the benchmark on which other readings can be based on. As it can be seen in Fig. 2a, the vibration of the idling vehicle on all axes is stable. The readings on the Z-axis is close to 1g due to the gravitational force. This reading is comparable to a smooth road as shown in Fig. 2d. In particular, the readings on all axes are relatively stable but with higher amplitudes than that of an idling vehicle.

Rumble strips are a series of raised strips that are typically painted across a road that causes the noise of a vehicle's tyres to change as it moves over them. These strips are also painted in high contrast colour such as yellow or white as a visual alert to drivers. In Fig. 2b, the readings from the Z-axis show multiple peaks followed by valleys, where each peak corresponds to a strip. The amplitude of the readings are also significantly higher than that of a smooth road, clearly distinguishing it from the latter. However, the Z-axis readings from the rumble strips are similar to the Z-axis readings on an uneven road as shown in Fig. 2e. Despite that, the uneven road can still be distinguished from the rumble strips due to the large fluctuations in the X and Y-axis as well. The strength of these fluctuations are directly proportional to the bumpiness of the road.

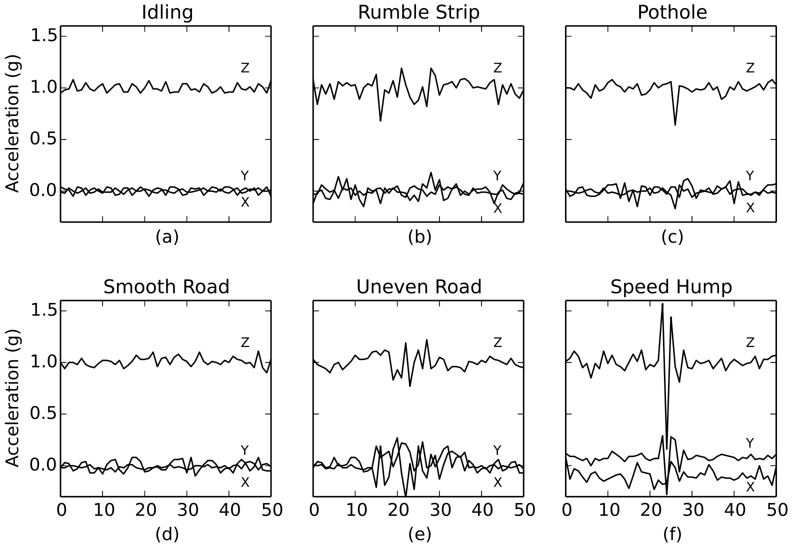


Fig. 2. Accelerometer readings for various road conditions; (a) engine running but vehicle is stationary, (b) rumble strip, (c) pothole, (d) smooth road, (e) uneven road, and (f) speed hump.

In Fig. 2c, the accelerometer readings indicate that the vehicle had driven over pothole. The single valley on the Z-axis shows a sudden downward motion due to the tyres falling into the pothole. Conversely, a speed hump is indicated by a peak on the Z-axis as shown in Fig. 2f. An interesting observation is that unlike a pothole where the readings return back to a neutral state ($\approx$ 1g) upon exiting the pothole, the speed hump causes a valley in the readings when the vehicle moves away from it. This valley is caused by the compression of the vehicle's suspensions, thus causing the vehicle to continue moving in a downward motion. The valley is immediately followed by another peak due to the vehicle's suspensions expanding and returning the vehicle back to the normal position.

As it can be seen from the plots, each of the road surface conditions can be clearly distinguished from the others. This clear distinction between them is advantageous when classification algorithms are used to automatically categorise each of these road conditions.

4.2 Performance Evaluation of Machine Learning Algorithms

The performance of the classifiers is measured using the Correctly Classified Rate (CCR). CCR is represented as the percentage of the correctly classified number of subjects divided by the total number of subjects in the dataset, as shown in Eq.1:

$$\mathrm{CCR} = \frac{T_P}{T} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

where TP is the True Positive rate, TN is the True Negative rate, FP is the False Positive rate, and FN is the False Negative rate. In its alternate form, T_P is the number of road surface conditions recognised and T is the total number of road surface condition data.

As with all machine learning-related experiments, we carried out the following two phases; training and testing. All data collected are used in both phases. For the **training phase**, the classification models are obtained by optimising each of the classifier's parameters. The most optimal combination of values for these parameters is determined empirically during the training phase. Some results of the empirically obtained parameters are shown in Fig. 3 while the actual values for these parameters are summarised in Table 2.

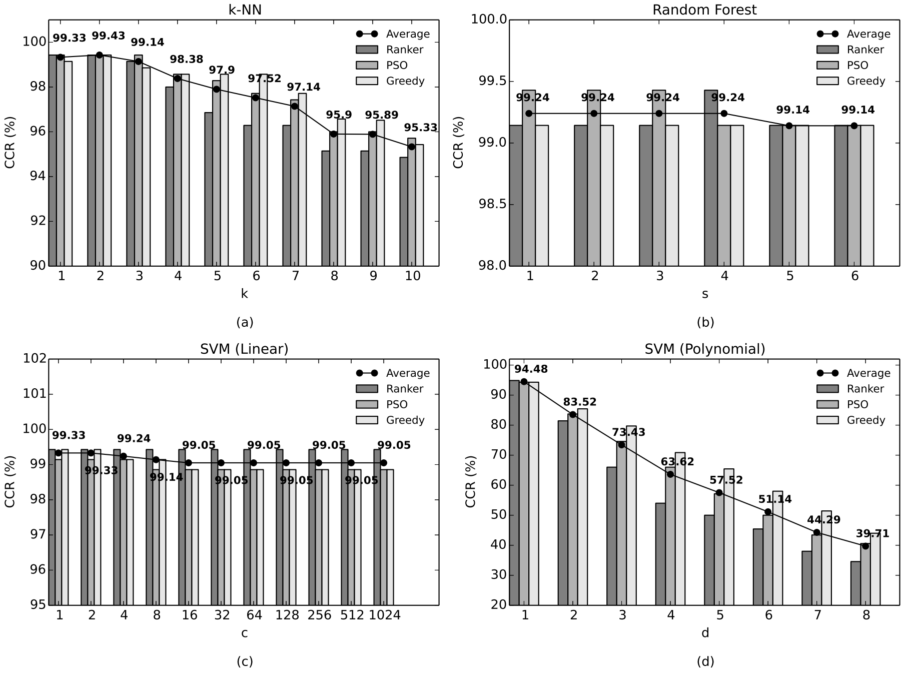


Fig. 3. Correctly Classified Rate for (a) k-NN, (b) Random Forest, (c) SVM (Linear), and (d) SVM (Polynomial) using classification models obtained from training phase.

In the subsequent **testing phase**, classification is performed using the models that we obtained from training. The feature selection algorithms and their

Table 2. Optimal parameter values for each classifier

Classifier	Optimal Parameters
k-Nearest Neighbour	$k = 2$
Random Forest	$s = 1$, $f = 3$, $i = 100$
Support Vector Machine (Linear)	$c = 1$
Support Vector Machine (Polynomial)	$c = 16$, $r = 0$, $d = 1$, $\gamma = 0.5$

Table 3. Comparison between different feature selection techniques and classifier

Classifier	CCR (%)			
	Ranker	PSO	Greedy	Average
k-NN	99.43	99.43	99.43	99.43
RF	99.14	99.43	99.14	99.24
SVM (Linear)	99.43	99.14	99.43	99.33
SVM (Poly)	94.86	94.29	94.29	94.48
Average	98.21	98.07	98.07	98.12

corresponding CCR for each of the four classifiers (including the different variations of SVM) are shown in Table 3. Based on the results shown in Table 3, all classifiers gave very high accuracy of greater than 99% except for SVM (Polynomial) which only had an accuracy of about 94%. Although not significantly poorer than the other classifiers, a possible reason why SVM (Polynomial) did not perform as well as the others is that the range of values for its parameters was not sufficiently optimal.

Generally, all three feature selection algorithms performed well in selecting the right features for classification. Some selection algorithms work better for some classifiers than others. For example, Ranker outperforms PSO when used together with k-NN and SVM (Linear) with a CCR of 99.43% but only scores a CCR of 99.14% for Random Forest. By comparison, PSO does better when used with Random Forest, with a CCR of 99.43%.

Future work to further improve the reliability and robustness of classifiers will see the addition of other features such as vehicle's linear acceleration and angular velocity calculated from the accelerometer readings using the Euclidean norm. Moreover, the frequency domain versions of these features can be included to mitigate the variations in the readings.

5 Conclusion

IoT sensors have been demonstrated to be both effective and also affordable when used as data collection devices to monitor the condition of road surfaces. The data from these sensors can then be channelled into machine learning algorithms to correctly classify the various road conditions. A classification model using statistical parameters is used for machine learning. Specifically, we tested three feature selection algorithms and another four classifiers to assess the performance of the proposed model. The model worked well in distinguishing and classifying the various road surfaces, with accuracy of more than 99%. This propose system will help facilitate the detection of poor road conditions, thus leading to faster repair and maintenance.

For our future work, we will consider the inclusion of additional features such as the vehicle's linear acceleration, angular velocity, and more frequency domain features. Additionally, the data from the IoT sensors can be pre-processed with appropriate filters to remove unwanted background noise such as inherent vehicle vibration and natural vibrations due smooth roads. This can potentially enhance the classification model. Overall, the proposed system has been shown to work well in the automated detection and classification of road surfaces.

Acknowledgements. Financial support from the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme with grant number FRGS/1/2015/SG07/MMU/02/1, as well as the Multimedia University Capex Fund with Project ID MMUI/CAPEX170008, are gratefully acknowledged.

References

1. Weka 3: Data mining software in java, https://www.cs.waikato.ac.nz/ml/weka/
2. Rm1.037 bil for federal road upkeep (June 2014), http://www.thesundaily.my/news/1069649
3. Two cousins, aged one and 11, killed in kota tinggi car crash (Nov 2016), https://www.nst.com.my/news/2016/11/192417/two-cousins-aged-one-and-11-killed-kota-tinggi-car-crash
4. Alqudah, Y.A., Sababha, B.H.: On the analysis of road surface conditions using embedded smartphone sensors. In: Information and Communication Systems (ICICS), 2017 8th International Conference on. pp. 177–181. IEEE (2017)
5. Astarita, V., Caruso, M.V., Danieli, G., Festa, D.C., Giofrè, V.P., Iuele, T., Vaiana, R.: A mobile application for road surface quality control: Uniqualroad. Procedia-Social and Behavioral Sciences **54**, 1135–1144 (2012)
6. Bhoraskar, R., Vankadhara, N., Raman, B., Kulkarni, P.: Wolverine: Traffic and road condition estimation using smartphone sensors. In: Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on. pp. 1–6. IEEE (2012)
7. Collotta, M., Pau, G., Salerno, V.M., Scatà, G.: A novel road monitoring approach using wireless sensor networks. In: Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on. pp. 376–381. IEEE (2012)
8. De Zoysa, K., Keppitiyagama, C.: Busnet–a sensor network built over a public transport system. Wireless Sensor Networks (2007)
9. Hull, B., Bychkovsky, V., Zhang, Y., Chen, K., Goraczko, M., Miu, A., Shih, E., Balakrishnan, H., Madden, S.: Cartel: a distributed mobile sensor computing system. In: Proceedings of the 4th international conference on Embedded networked sensor systems. pp. 125–138. ACM (2006)

10. Mednis, A., Elsts, A., Selavo, L.: Embedded solution for road condition monitoring using vehicular sensor networks. In: Application of Information and Communication Technologies (AICT), 2012 6th International Conference on. pp. 1–5. IEEE (2012)
11. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In: Proceedings of the 6th ACM conference on Embedded network sensor systems. pp. 323–336. ACM (2008)
12. Perttunen, M., Mazhelis, O., Cong, F., Kauppila, M., Leppänen, T., Kantola, J., Collin, J., Pirttikangas, S., Haverinen, J., Ristaniemi, T., et al.: Distributed road surface condition monitoring using mobile phones. In: International Conference on Ubiquitous Intelligence and Computing. pp. 64–78. Springer (2011)
13. Sun, X.g., Sun, X.l., Yang, Q.g., Ma, S.n.: Application of wireless sensor networks in post-disaster road monitoring system. In: Intelligent Networks and Intelligent Systems (ICINIS), 2011 4th International Conference on. pp. 105–108. IEEE (2011)

Real-Time Optimal Trajectory Correction (ROTC) for Autonomous Omnidirectional Robot

Noorfadzli Abdul Razak, Nor Hashim Mohd Arshad, Ramli bin Adnan, Norashikin M.Thamrin, Ng Kok Mun.

Universiti Teknologi Mara , Shah Alam , Selangor 40450 , Malaysia
noorfadzli@gmail.com

Abstract. This paper proposed a Real-Time Optimal Trajectory Correction (ROTC) algorithm designed to be applied for autonomous omnidirectional robot. It is programmed to work when a robot undergoes a deviation, an admissible trajectory correction path is generated for the robot rapidly returns to the route line. For the algorithm to do this, initially a deviation scheme is employed to sense deviation and formulates a vector consists of displacement and angle. Via the vector, an admissible correction path is originated utilizing Hermite cubic spline method fused with time and tangent transformation schemes. A Dead Reckoning (DR) technique is applied for robot to pursue the path. Several experiments are arranged to evaluate the reliability of robot navigation with and without the algorithm. It motion is mapped in Graphical User Interface (GUI) window using data from Laser Range Finder (LRF) sensors as attached to the robot controller. Using the map, the performances of the algorithm are evaluated in terms of distance travel and duration to return on the line. The results signify robot navigation with the algorithm required shorter distance and duration as compared to robot navigation without ROTC. Thus, it justifies the algorithm is feasible in the navigation system where it can assist robot effectively to move back to the route line after experiencing a deviation caused by a disturbance.

Keywords: Deviation, Line Tracking Technique, Autonomous Control, Omnidirectional Robot; Trajectory Correction Path.

1 Introduction

Applying omnidirectional robot to handle logistic tasks such as transport, sorting, delivery of products and others in industry becomes relevant nowadays. Implementing autonomous navigation such as line tracking technique either using several photo sensors or a camera increases the robot usability to perform the tasks, especially in tight areas. However, there is a major drawback encountered by this robot where when it diverges from route line due to disturbance, it suffers difficulty to navigate back to the line. The disturbance in this situation is referred to false reading from sensors or camera when tracking the line or may result from robot motion itself performing obstacle avoidance. There are several researchers did propose a corrective motion technique which can be employed to enable the robot to return to the route

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_27

line. Such as Seiler et al [1] proposed a motion technique using Lie group of symmetries. Pham et al [2] presents corrective motion algorithm based on affine transformation. For Sprunk et al [3], a kino-dynamic trajectory generation technique is suggested. Meanwhile, Künemun et al [4] designed a fast and accurate generation of cubic spline trajectories.

As for this research, their works become an inspiration for us to introduce a new technique that able to perform the similar action. Hence, a novel algorithm name as ROTC is presented. This algorithm is designed for situation when the robot diverges from route line because of disturbance, then it will generate an admissible trajectory path for the robot to pursue in order to return to the line. Admissible path refers a closest path to the line and can be pursued by robot where it must not exceed its maximum velocity. Deviation vector comprises of distance, and angle is used by the algorithm to yield the said path. It will be a cubic type as applied by Zhou et al [5] and Hermite interpolation technique is fused in the algorithm to yield it. This type of path is selected since it can preserve the robot navigation momentum and evades backlash to the drive mechanisms. Meanwhile, using Hermite technique allows the path easily can be produced using two points and tangent vectors. Special transformation schemes namely, tangent and time are integrated in the algorithm. For a tangent transformation scheme, it used to alter the target tangent vector so that the correction path becomes near to route line. In the meantime, time transformation scheme will ensure velocity of each waypoint on the path can be pursued by the robot. Once the admissible path is obtained, a DR technique is employed to move the robot to each waypoint on the path. The robot is tested in three deviation acceleration categories to justify the algorithm dynamic functionality. Comparative performance is carried where robot navigation with and without the ROTC algorithm are evaluated.

The process implemented to develop and evaluates the ROTC algorithm performance are described throughout this paper according to arrangement as follows. The workflow of the ROTC algorithm is explained in Section 2. Next, methodologies perform in the research are particularized in Section 3. Section 4 elaborates the experiment setup, and approach arranges to evaluate the ROTC performance. In Section 5, it will present and discuss outcomes gained from the experiments. At last, Section 6 concludes the overall works in this research.

2 Design Concept

2.1 ROTC Workflow

Fig.1 demonstrates the pictorial diagram to illustrate the ROTC algorithm workflow to generate an admissible trajectory correction path when an autonomous robot experiences a deviation. According to Fig.1, an omnidirectional robot equipped with line tracking technique via a camera is arranged to navigate and pursuing the line. In the event, a disturbance causes the robot to diverge from the route; a deviation vector is estimated using an approach as introduced by Razak et al [6]. At that moment, ROTC algorithm is activated and begins to generate an initial linear trajectory path based on deviation vector, $\overrightarrow{\sigma}$that comprises displacement, s_σ and angle θ_σ. Then, there are two scheme are designed namely the tangent and time transformations to correct the path to become admissible.

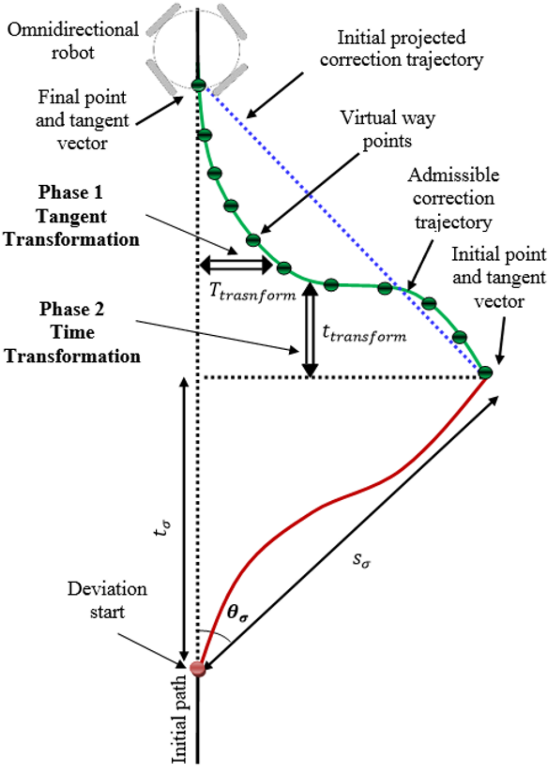

Omnidirectional robot
Initial projected correction trajectory
Final point and tangent vector
Virtual way points
Admissible correction trajectory
Phase 1 Tangent Transformation
Initial point and tangent vector
$T_{trasnform}$
Phase 2 Time Transformation
$t_{transform}$
t_σ
s_σ
Deviation start
θ_σ
Initial path

Fig. 1. ROTC workflow

In the tangent transformation phase, the algorithm will adaptively adjust the trajectory final tangent vector. The tangent is varied until the resulting trajectory achieves maximum proximity to the original route. At the same time, the series of points along the trajectory must be ensured ascending to each other. This is for attaining the continuity and smooth transitions throughout the vehicle maneuvers. Once the first scheme establishes the correct tangent for the trajectory, the second scheme which involves time transformation will commence. The average velocity at each segment of two points along the trajectory is computed. Then it will be validated with maximum velocity that allowable to be enforced by the robot. This maximum velocity is theoretically calculated based on the specification of motors employed in the vehicle system. If one of the segments has velocity exceeds the limitation, therefore, the trajectory yielded from first phase becomes invalid. Hence, automatically the algorithm makes an adjustment on the time navigation. At this stage, the time navigation will be increased. Simultaneously, the first scheme is activated again. This process will recurrently be executed until both schemes able to identify the best final tangent vector and optimal navigation time for the trajectory to lead back the vehicle to original route. At last, the trajectory correction that is expressed as admissible produced. The average velocity at each segment of the trajectory is also secured. Alongside, via the velocity, the navigation course or angle at the segment can be formulated. Hence, with the data, it is now conceivable to activate following framework sequence namely DR navigation technique. This technique is chosen because the technique is simple and practicable enough for directing the robot to pursue on the generated path. Therefore, the robot will be moved until the final point of trajectory. At the same time, once the camera able to sense the route line back, then navigation via line tracking is activated again, and all the RATC sequences will be reset. The mentioned sequences and process will continuously run every time the robot experienced a deviation caused by the disturbance. This is also applicable either deviation happened to the left or right from the route line.

3 Methodology

This section reveals the process to design the proposed ROTC algorithm. It includes on how the trajectory is developed using a cubic Hermite interpolation method along with respective equations. In addition, the ways the cubic trajectory path is corrected to be admissible by tangent and time transformations are clarified. Then, a custom-made controller utilized to implement the algorithm, line tracking and DR navigation techniques is exposed. This goes along with devices installed on it. Afterward, the section confers a customize robot as an experimental platform, and its control model. At last, a GUI window employed for data acquisition and maps robot motion are presented.

3.1 ROTC algorithm

Once the algorithm activated, it begins to project a linear virtual path, σ_{lin}. This path is projected inversely relative to last location of vehicle deviation toward to original navigation route. In time, this location becomes an initial navigation point, p_{init} for the vehicle to maneuver back on the route. The displacement, s_{lin} of the σlin have similar values of s_σ as recorded by the deviation scheme. The same thing also happened to the angle, θ_{lin} of the σ_{lin}, where the amount is set equally to deviation angle, θ_σ. However, the direction of the angle at this moment is overturned. The σ_{lin} is deliberately yielded to fix the 2D maximum permissible position, p_{max} on the line route for the vehicle to enter. Likewise, this position will become the target location, p_{trgt} for the vehicle to pursue. Therefore, it is essential to determine the target location, in XY coordinate system and listed are formulas utilized.

$$p_{y\_trgt} = s_\sigma \sin\theta_\sigma \tag{1}$$

$$p_{x\_trgt} = s_\sigma \cos\theta_\sigma \tag{2}$$

Hence, p_{init} and p_{trgt} for robot navigation in the 2D coordinate system at this instant are recognized. Next, the action will be executed by the algorithm is formulating an admissible trajectory correction path. To do this, since the research does fix the path will be in form of cubic, therefore, it equations in 2D coordinate as follows:

$$p_y(t) = a_{y3}t^3 + a_{y2}t^2 + a_{y1}t + a_{y0} \tag{3}$$

$$p_x(t) = a_{x3}t^3 + a_{x2}t^2 + a_{x1}t + a_{x0} \tag{4}$$

The t in the equations is concerned with the time variable. Meanwhile, a_3, a_2, a_1 and a_0 are the coefficients for the equations. These coefficients must be resolved since they will influence the form of the path. Coefficients for $p_y(t)$ will be resolved first and at the same time, similar ways are implemented to find coefficients for $p_x(t)$. First, a_0 in Equation 1 will be identified. In Fig.1, it is shown that the initial point of the trajectory path is p_{init}. Correspondingly, the $p_y(t)$ in Equation 1 turns out to be $p_{y\_init}(t)$, where the value is set to be zero. Furthermore, at that point, t is 0 second. Hence, substituting the $p_{y\_init}(t)$ and t in Equation 3 causes a_{y0} established as listed below:

$$a_{y0} = p_{y\_init}(0) = 0 \tag{5}$$

Next, to find a_{y1}, Equation 1 is differentiated and result in following equation is obtained.

$$p_y{}'(t) = 3a_{y3}t^2 + 2a_{y2}t^1 + a_{y1} \tag{6}$$

This equation actually presents a tangent vector denoted as m_x for the path in X-axis domain. Via this equation, the a_{x1} now can be identified via following means. The initial tangent vector, m_{y_init} is formulated by substituting t as 0. This is done because it is located at p_{init}. As a result, a_{y1} becomes as shown:

$$a_{y1} = p_y{}'(0) = m_{y\_init} \tag{7}$$

To this extent, only a_{y2} and a_{y3} are still unresolved. In order to determine them, it is necessary to obtain the target point, $p_{y\_trgt}(t)$ and target tangent vector, m_{y_trgt} equations. To do that, it is assumed the robot arrived at *ptrgt,* at 1 second, which make t can be set as 1. Therefore, substituting the t along with a_{y1} obtained previously in Equations 3 and 6 give the desired equations appear as follows:

$$p_{y\_trgt}(1) = a_{y3} + a_{y2} + p_y{}'(0) \tag{8}$$

$$m_{y\_trgt} = p_y{}'(1) = 3a_{y3} + 2a_{y2} + p_y{}'(0) \tag{9}$$

Noticeably, Equations 8 and 9 turn to be in form of linear equations. This permits matrix manipulation method can be employed to find a_{y2} and a_{y3}. Applying the said method gives both coefficients are recognized as stated:

$$a_{y2} = 3p_{y\_trgt}(1) - 2p_y{}'(0) - p_y{}'(1) \tag{10}$$

$$a_{y3} = -2p_{y\_trgt}(1) + p_y{}'(0) + p_y{}'(1) \tag{11}$$

All the coefficients at present are established, hence substituting them in Equation 3 results in the cubic trajectory path in Y-axis domain can be finalized as follows:

$$p_y(t) = \begin{bmatrix} t^3 \\ t^2 \\ t^1 \\ t^0 \end{bmatrix}^T \begin{bmatrix} -2p_{y\_trgt}(1) + p_y{}'(0) + \ p_y{}'(1) \\ 3p_{y\_trgt}(1) - \ 2p_y{}'(0) - p_y{}'(1) \\ p_y{}'(0) \\ 0 \end{bmatrix} \tag{12}$$

Meanwhile, similar approaches above are utilized to recognize coefficients for trajectory path in X-axis domain. Therefore, this path equation is presented in Equation 13.

$$p_x(t) = \begin{bmatrix} t^3 \\ t^2 \\ t^1 \\ t^0 \end{bmatrix}^T \begin{bmatrix} -2p_{x\_trgt}(1) + p_x{}'(0) + \ p_x{}'(1) \\ 3p_{x\_trgt}(1) - \ 2p_x{}'(0) - p_x{}'(1) \\ p_x{}'(0) \\ 0 \end{bmatrix} \tag{13}$$

At this level, the path equations in Y and X axis domains with respect to *t* are obtained. Now, it is organized; these paths comprise numbers of waypoints labeled as *n*. Correspondingly, each *n* can be computed via these equations by setting the time as $t \in \left(\frac{t_{nav}}{10}\right)n$ where n $\in \{0,1,2 \ \cdots, 10\}$. Giving that the p_{init}, p_{trgt} and m_{init} are known, hence leaving m_{trgt} becomes indefinite. In concern about this matter, the algorithm has introduced tangent and time transformation schemes to find the correct for m_{trgt}. This is to ensure the path generated will near to line route and can be pursued by robot at optimum velocity. The flowchart in Fig.2 demonstrates the workflow of the schemes.

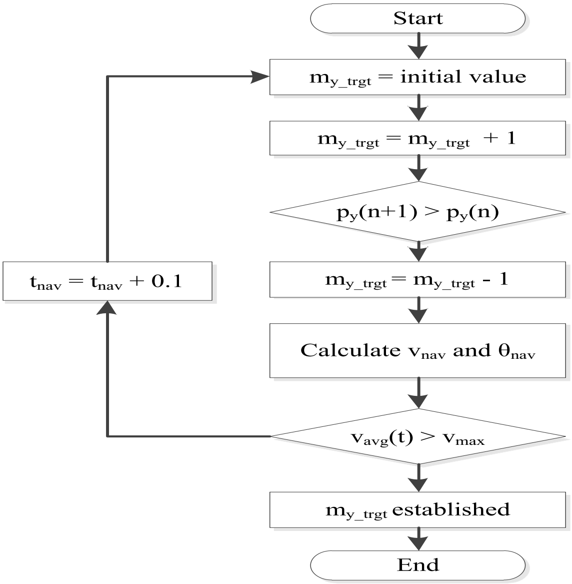

Start
m_{y_trgt} = initial value
$m_{y\_trgt} = m_{y\_trgt} + 1$
$p_y(n+1) > p_y(n)$
$m_{y\_trgt} = m_{y\_trgt} - 1$
Calculate v_{nav} and θ_{nav}
$v_{avg}(t) > v_{max}$
m_{y_trgt} established
End
$t_{nav} = t_{nav} + 0.1$

Fig. 2. Workflow of tangent and time transformation scheme

In the tangent transformation scheme, a preliminary value for m_{trgt_y} is assigned. Then it will be elevated one causes the path form to change. At moment, each position of n on the path will be reviewed. The review process must obey the rule where every post point must be more than prior point. If the rule obeyed, than m_{trgt_y} will be increased one. The process continuously runs until the rule is defied. Once happened, the current m_{trgt_y} value will be minus by one since the previous value is actually the optimum m_{trgt_y} that causes the path nearer to line. Next, time transformation scheme is activated. The average velocity, v_{avg} between each segment of waypoints on the X and Y axis paths are calculated using Equation 14 and 15

$$v_{y_{avg}}(t) = \frac{p_{y[n+1]}\,(t+0.1)-\,p_{yn}\,(t)}{(t+0.1)-t} \tag{14}$$

$$v_{x_{avg}}(t) = \frac{p_{x[n+1]}\,(t+0.1)-\,p_{xn}\,(t)}{(t+0.1)-t} \tag{15}$$

Respectively, the average velocity for each segment on 2D path and also signified as navigation velocity, v_{nav} can be obtained with subsequent equation:

$$v_{nav}(t) = \sqrt{\left(v_{xn\_avg}\,(t)\right)^2 + \left(v_{yn\_avg}\,(t)\right)^2} \tag{16}$$

Meanwhile, the navigation angle, θ_{nav} of the segments are also computed via following equation

$$\theta_{nav}(t) = tan^{-1}[\frac{v_{yn_{avg}}(t)}{v_{xn_{avg}}(t)}] \tag{17}$$

Next the scheme will check either the v_{avg} of each segment is less than v_{max}. If yes, then the m_{trgt_y} value is established and the path generated is optimum for robot to return to the lone. Contrary, if not, subsequently the time is increased by 0.1s. Tangent transformation scheme is started again, and this process continuously runs until m_{trgt_y} value that ensures velocity at each segment less than v_{max} is gained.

3.2 Robot control model and navigation techniques

A custom-made omnidirectional robot using four wheels is built to be employed as an experiment platform. Acrylic sheet and aluminum bar is used as robot chassis. The chassis has dimension of 30 cm height and 60 cm for it width and long. The robot is driven using four 7.2V DC motors and every motor coupled with 2 inches transwheel. Overall robot weight is 3.2 kg. Meanwhile to enable the robot to move according to navigation inputs namely v_{nav} and θ_{nav}, a control model for the robot is required. To obtain the model, Fig.3 shows a free body diagram of the omnidirectional robot. According to the diagram, symbol O as located at the chassis center is meant for robot gravity. Thus, robot local coordinate in X and Y axis becomes X_LOY_L. As for robot global coordinate, it turns to be X_GOY_G. The α_i is angle between the axles of the wheels relative to X_LOY_L. Character i is related to the number of wheel used by the robot. Since the robot used four wheels hence it is fixed to $\alpha_1 = 90^0$, $\alpha_2 = 180^0$, $\alpha_3 =$

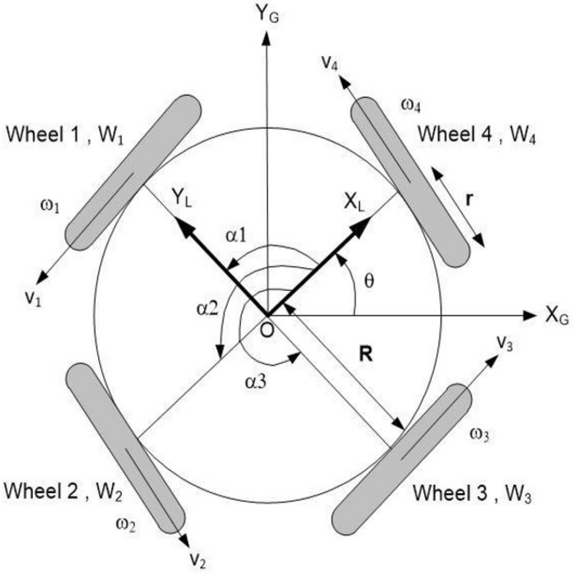


Fig. 3. Omnidirectional robot free body diagram

270^0 and $\alpha_4 = 0^0$. The ω_i is denoted for the angular velocity of each wheel, and v_i signifies the direction of linear velocity of the center of the wheel with respect to X_LOY_L. This robot has a chassis radius, R of 30 cm and radius of the wheel, r is 2.54 cm. Meanwhile, θ appears to be angle between the coordinates X_LOY_L and X_GOY_G. Hence, via the diagram and applying the methods as carried out by Alves et al [7], the control model for the robot is like so:

$$\begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \end{bmatrix} = \frac{1}{r} \begin{bmatrix} -\sin(\theta + \propto_1) & \cos(\theta + \propto_1) & R \\ -\sin(\theta + \propto_2) & \cos(\theta + \propto_2) & R \\ -\sin(\theta + \propto_3) & \cos(\theta + \propto_3) & R \\ -\sin(\theta + \propto_4) & \cos(\theta + \propto_4) & R \end{bmatrix} \begin{bmatrix} V_{Gx} \\ V_{Gy} \\ \dot{\theta} \end{bmatrix} \tag{18}$$

In this model, the V_{Gx} and V_{Gy} represent velocities of X and Y axis in X_GOY_G. However, to appropriate steers the robot, it is best to translate V_{Gx} and V_{Gy} into velocities in X_LOY_L. This can be done via Equation 19 to 21.

$$\begin{bmatrix} V_{Gx} \\ V_{Gy} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & 0 & 0 \\ 0 & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_{Lx} \\ V_{Ly} \\ \dot{\theta} \end{bmatrix} \tag{19}$$

where:

$$V_{Lx} = V_{nav} \cos \theta_{nav} \tag{20}$$

$$V_{Ly} = V_{nav} \sin \theta_{nav} \tag{21}$$

With these equations, it is possible for the model to receive v_{nav} and θ_{nav} and then determine correct ω_i so that robot can move to desired direction and velocity. In the meantime, the robot is arranged to perform two navigation techniques line tracking and Dead Reckoning (DR). To enable the robot to navigate via line tracking technique, an analog camera is utilized. Camera will capture the line image and then a selective video line technique designed by Arshad et al [8] to track the line. The technique will produce digital values represent the location of the line. Via the values, robot is programmed to move at constant v_{nav} and θ_{nav} is regulated so that robot steers in proportion to the line. As for navigation via DR technique, it is used to steer the robot to move along the correction path. Each segment of waypoints on the path has it own navigation inputs namely v_{nav} and θ_{nav} as calculated via Equations 16 and 17. Therefore, a timer interrupt is exploited by the controller to supply these inputs to the control model according to interval t. By doing this robot is ensured to perform the desired motion to follow the path.

3.3 Controller board and GUI window

To execute the ROTC algorithm and autonomous navigation of the robot, a customize controller board based on PIC32MX795F512L microcontroller is built. The board is equipped with an accelerometer, a camera, four units of 10A motor driver, two units of LRF sensors and a XBee 2.4GHz module. All of these devices are needed so that the desired functions can be performed. Like an accelerometer, it is meant for deviation scheme to compute the vector. A camera is employed for line tracking technique. Motor driver is used to drive each motor according to respective ω_i . Temporarily, the LRF sensors are used to provide 2D distance data. These data are then transmitted wirelessly to a custom-made GUI window via the XBee device. The GUI window is developed using Microsoft Visual Basic 2015 and installed in a PC. A same XBee module installed on the controller board is plugged to a PC. Hence, the distance data received are exploited to map the robot motion. The motion includes robot motion when autonomously pursuing the line route, deviations that happened and navigates on the trajectory correction path to return to the route. Meanwhile, the controller and robot itself are powered by 11.1V, 2.2A Lithium Polymer (Li-Po) battery.

4 Experimental setup and data validation

An indoor controlled environment field as shown in Fig.4 is utilized to test the proposed algorithm performance. The field is built using a grey rubber mat where the dimension is 6 meters long and 3 meter width. Additionally, a white line with 3cm width is taped at the center of the mat. With this setup, it is possible for a camera to track the line and enable the robot to perform autonomous navigation. Meanwhile, there are few wood blocks positioned within the field. They are utilized as object to reflect laser beams from LRF sensors to gain 2D distance data. These data are necessary to map robot motion trail when moves along the line. Consequently, the map is used for evaluation purpose afterward.

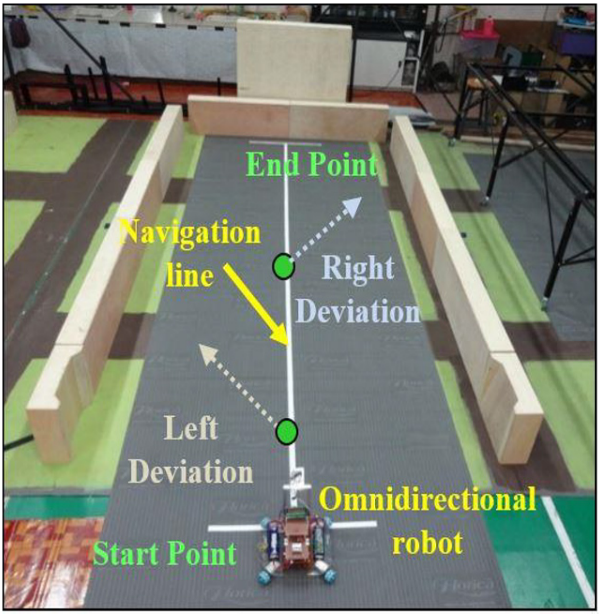


Fig. 4. Experiment field

The trials begin by positioning the robot at the start point of navigation line. Then, robot is activated to move at a constant speed. At a distance from the start point, the robot is pulled via a string to make it deviate to the left from the route line. Doing this action actually simulates the same effect when the robot suffers deviation due to disturbance in real practice. At that moment, the process where line tracking technique is disabled, a deviation scheme will sense deviation and computes the vector; ROTC generates the path and finally, DR technique moves the robot to return to the line are monitored. Concurrently, the respective data are captured in GUI window. As the robot able to return to the line, it is given a time to move steadily on the line. Then the string is pulled once more to force the robot to diverge to the right from the line path. Similar robot process as mentioned above is monitored again and data are captured in GUI window. The experiment is stopped once the robot reaches the end position on the line route.

Subsequently, the experiment is performed again whereas at this time, robot autonomous navigation without the algorithm is implemented. The experiment is not fully executed as applied in experiment before. The vehicle will be moved manually at end position of deviation. The position is easily tracked by using LRF modules. The navigation path before and during deviation occurred in prior experiment is presumed to be similar for this navigation. This goes along with the duration of the navigation. Next, the vehicle is activated again to move and senses the route via line tracking technique. Simultaneously, navigation time is resumed and the GUI maps the navigation path. The vehicle is programmed to stop once the camera able to track the line of route. Once it stops, the motion path is unified with the navigation path as recorded earlier. As for the duration of the navigation, it will be added to time as recorded before. In the meantime, these steps will be repeated for the situation of second deviation that occurred toward the right of the route.

Accordingly, the methods as explained above is decided since the methods are the best way to observe and compare the path navigated by vehicle with and without ROTC algorithm. Particularly, to ensure both navigation systems experience the same deviation distance and direction. Meanwhile, this experiment has been set up to run in three different categories of deviation acceleration. The categories are high, moderate and low having the range of acceleration as specified in Table 1. Such setup is arranged to test the dynamic operation of the designed algorithm to perform in different acceleration of deviation. To validate the algorithm performance, the map that plots robot navigation with and without the algorithm is examined. Via the map, analytical evaluation can be carried out to differentiate performance of robot navigation with and without the algorithm in terms of distance and time taken to return to the original path after experiencing deviation.

Table 1. Deviation acceleration categories

Category	Deviation acceleration , a (ms^{-2})
High	$a > 10$
Moderate	$5 \leq a \leq 10$
Low	$a < 5$

5 Results and Discussion

Fig. 5 to Fig. 7 show the maps that plots robot motion either it autonomous navigation is integrated with and without ROTC algorithm in all deviation acceleration categories. Apparently as seen in the figures, robot with autonomous navigation equipped with ROTC algorithm performs well when maneuvers on the route in all the categories. Even the robot endures deviation either to the left or right, somehow the algorithm guarantees the robot to return nearly to the original line Moreover, it also can be realized, the distance taken for the robot where it navigation fused with ROTC algorithm appears to be shorter. Consequently, robot consumes lesser time to return to the line. This achievement is compared to robot navigation without the algorithm where the distance and duration turn out to be longer. All of above declarations actually justifies decisively according to data as summarized in Table 2. Meanwhile, the table also reveals the performance of robot navigation with the algorithm becomes deteriorated from high to medium and subsequently to low category. The duration for the robot to return to the line appears to increase slightly from one category to another. This goes along with the distance traveled by the robot. This matter happened because of inaccuracy of deviation scheme to estimate the vector. According to Razak et al [6], the accuracy of the deviation scheme that designed by them becomes declined as the deviation acceleration decrease. Due to this drawback, the correction path generated by the ROTC algorithm actually doesn't reach the line route.

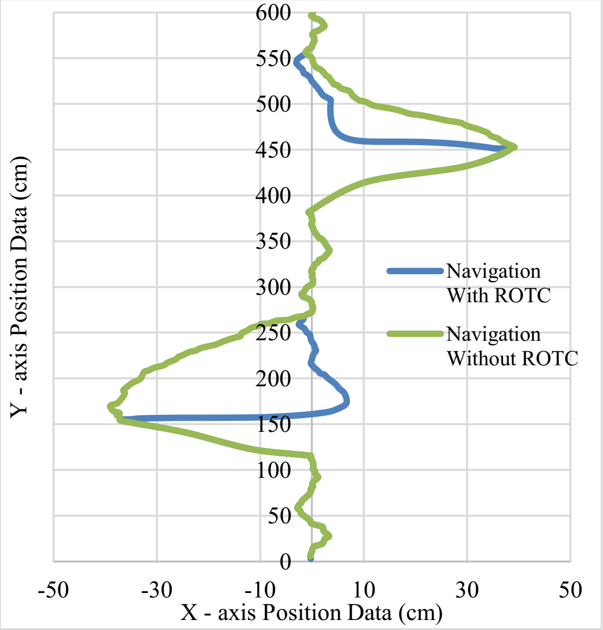


Fig. 5. Robot motion in high acceleration category

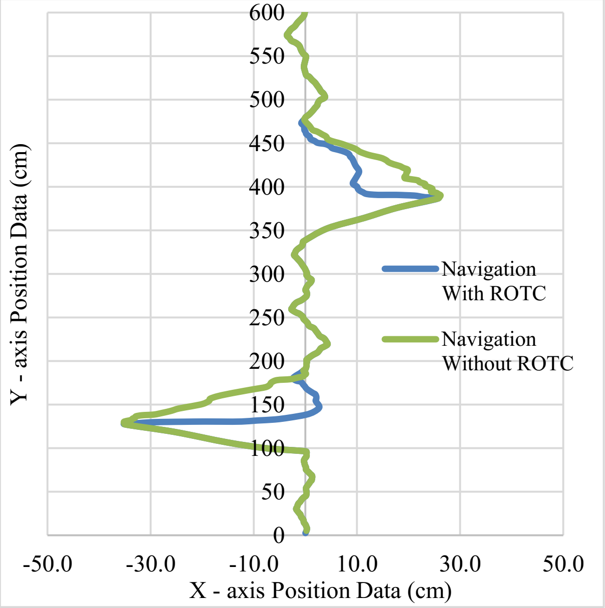


Fig. 6. Robot motion in moderate acceleration category

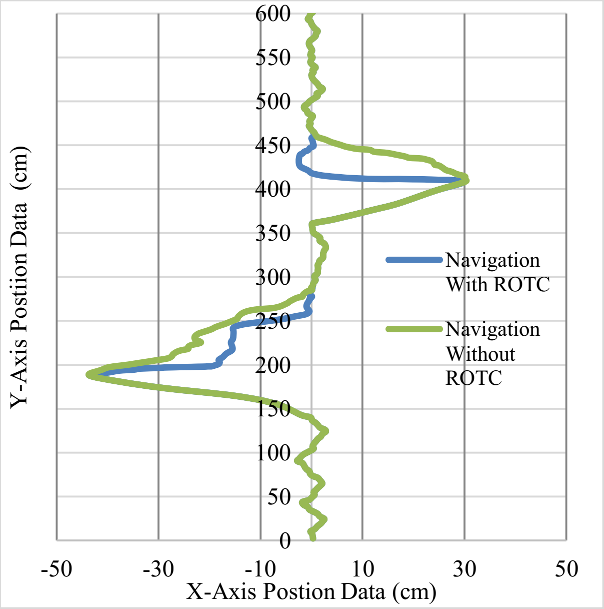


Fig. 7. Robot motion in low acceleration category

Table 2. Results for robot navigation with and without ROTC

Deviation		Navigation Time (s)		Navigation Distance (cm)	
Category	Direction	With ROTC	Without ROTC	With ROTC	Without ROTC
High	Left	2.1	2.7	81.0	89.3
	Right	2.5	3.0	79.5	82.5
Moderate	Left	2.9	4.3	85.3	95.4
	Right	3.3	4.0	87.3	91.4
Low	Left	3.8	4.8	94.0	110.3
	Right	3.7	4.7	97.4	113.5

This causes when the robot arrives at end location of the correction path, more time is needed for the camera to track the line. Therefore, robot navigates further to reach the line again and result in the duration turn out to be increased. Nevertheless, at moment, the research does justify with these achievements, the usability of the ROTC algorithm to assist robot navigation to return to lined route after experienced is practically relevant.

6 Conclusion

A ROTC algorithm designed using Hermite Spline Interpolation technique is established. Integrated with tangent and time transformation schemes enables it to generate an admissible trajectory path. Using DR technique, robot can pursue the path to return rapidly to the lined route after experienced deviation due to disturbance. Three categories of deviation acceleration are carried out to test algorithm performance applied to an autonomous omnidirectional robot. It performances are compared with robot navigation without the algorithm. The outcomes do reveal, the dynamic practicality of the algorithm to work in different deviation acceleration. Moreover, the results prove robot navigation with ROTC performs better where the distance and duration for robot to return to route line shorter compared to robot navigation without it. Even there is performance deterioration of the robot when experimented in high, medium and then low category, however, this achievement is accepted since via the algorithm robot still able to return the lined route.

Acknowledgment

This research is supported by Ministry of Education fund (600-RMI/RAGS 5/3 (42/2015)). The authors would like to express appreciation to Faculty of Electrical Engineering and Institute of Research Management and Innovation, Universiti Teknologi Mara (UiTM) for supplying the financial supports and equipment to conduct this research.

References

1. Seiler, K. M., Singh, S. P., Sukkarieh, S., Durrant-Whyte, H.: Using Lie group symmetries for fast corrective motion planning. The International Journal of Robotics Research 31(2), pp. 1-16 (2012).
2. Pham, Q. C.: Fast trajectory correction for nonholonomic mobile robots using affine transformations. In Robotics: Science and Systems VII, pp. 265-272 (2012).
3. Sprunk, C., Lau, B., Pfaffz, P., Burgard, W.: Online generation of kinodynamic trajectories for non-circular omnidirectional robots. In IEEE International Conference on Robotics and Automation (ICRA), pp. 72-77 (2011).
4. Künemund, F., Kirsch, C., Heß, D., Röhrig, C.: Fast and accurate trajectory generation for non-circular omnidirectional robots in industrial applications. In 7th German Conference on Proceedings of ROBOTIK, pp. 1-6 (2012).
5. Zhou, F., Song, B., Tian, G.: Bezier Curve Based Smooth Path Planning for Mobile Robot. Journal of Information &Computational Science 8(12), pp. 2441-2450 (2011).
6. Razak, N. A., Arshad, N. H. M., Adnan, R., Misnan, M. F., Thamrin, N. M., Mahmud, S. F.: A Real-Time deviation detection and vector measurement technique for straight line quadrocopter navigation using accelerometer. In IEEE 5th Control and System Graduate Research Colloquium (ICSGRC), pp. 69-74 (2014).
7. Alves, S. F., Rosario, J. M., Ferasoli Filho, H., Rincon, L. K., Yamasaki, R. A.: Conceptual bases of robot navigation modeling, control and applications. In Advances in Robot Navigation (2011).
8.Arshad, N. M., & Razak, N. A.: Vision-based detection technique for effective line-tracking autonomous vehicle. In IEEE 8th International Colloquium on Signal Processing and its Applications (CSPA), pp. 441-445 (2012).

Incorporating Cellular Automaton based Microscopic Pedestrian Simulation and Genetic Algorithm for Spatial Layout Design Optimization

Najihah Ibrahim[1], Fadratul Hafinaz Hassan[1] and Safial Aqbar Zakaria[2]

[1] School of Computer Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
[2] School of Housing, Building and Planning, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
najihah.ibrahim@student.usm.my, fadratul@usm.my, [3]ssafial@usm.my

Abstract. Autonomous spatial layout design had become the prominent applications for early planning process for a space arrangement. The available spatial layout design applications had focused on the structural and functional demands for a trendy, cultural influences and scalable spatial design. The previous research on the autonomous spatial layout design had proved that the Genetic Algorithm (GA) with full-fledged of operators are able to design an optimal spatial layout that is scalable for the space utilization. However, in these recent years, due to many occurrences of emergency incidents, the security assurance of the pedestrian to evacuate from the spatial layout during panic situation had become the main focus in designing a spatial layout. Hence, this research had proposed the pedestrian movement simulation using Cellular Automata (CA) with the Moore Neighborhood transition movement direction as the objective function for optimizing the GA operators based spatial layout design. The results have shown that the CA based pedestrian movement simulation was able to optimize the GA operators based spatial layout design in designing a feasible and scalable space arrangement with low-risk of pedestrian casualties.

Keywords: Spatial Layout Design, Genetic Algorithm, Pedestrian Simulation, Cellular Automata, Optimization.

1 Introduction

Designing a space by using an autonomous spatial layout design application had become the most common practice in nowadays interior design planning process [1-3]. In the early time, spatial layout design was designed by using the manual technique that focused on the standard structural design with the influence of the local demographic structure's culture [4, 5]. However, technologies had made a great change on the computer exploitation for the autonomous spatial layout design with a great design's quality, as good as the manual design's marker with easier and faster in designing process and providing a great layout [6-8]. As the autonomous layout design become the new ubiquitous subject in constructing a structural design for a space, the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_28

functionality and space utilization had become the next focus point for enhancing the spatial layout design to comfort and equip the human's daily routines and needs [9].

Nowadays, there are several available spatial layout design tools that were developed and easily can be operated while handy in assisting the user for designing the spatial layout with high functionality and optimal space utilization [6-8]. In previous research, the GA based spatial layout design with full-fledged of operators (selection, crossover and mutation) was constructed and able to find the feasible layout with optimal space utilization [10]. However, the structured spatial layout with fully fledged of space's functionality and scalable space utilization will not be enough in constructing a spatial layout design as one of the other factor in designing a spatial layout design is the crowd-traffic movement [11-13].

In this recent years, there are a lot of incidents happened that leads towards high casualties due to the bottleneck effect of the pedestrian movement. In Malaysia, there were two major incidents that had caused great fatalities of the pedestrian; 1) A religious school burned down and 2) Total death of family of four in a great fire [14, 15]. There are also some major incidents that happened due to the bottleneck effect and high physical collision between the pedestrian such as; the Mina stampede at Mecca in 2015 during hajj pilgrims, the Station Nightclub stampede at Rhode Island in 2003 due to the sudden fire combustion and the crowd overflow at the Address Downtown Hotel in Dubai during new year's eve celebration [16, 17]. These incidents happened due to the difficulty to escape and the time consuming route plan for the victim to evacuate and move towards the exit point within short amount of time. Hence, the victims involved were entrapped inside the building and were burned to death. Due to these incidents, the process of designing a spatial layout had become more crucial as the crowd traffic activities had become one of the focal point in space arrangement besides the security, functionality, economic, attraction, cultural, structural and many other purposes. Fig. 1 shows the overview of basic elements on designing the autonomous spatial layout design.

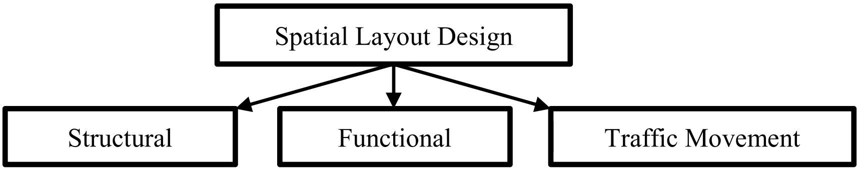


Fig. 1. Overview of spatial layout design parameter in designing the autonomous spatial layout design

Based on Fig. 1, the pedestrian movement had been shown as one of the objective function besides the structural and functionality for a spatial layout design [11, 12]. This Fig. 1 is supported by the Lewin's Equation;

$$\mathrm{B} = f(\mathrm{P},\mathrm{E}) \tag{1}$$

Equation 1 shows the Lewin's Equation where **B** is the behavior or reaction, **P** is the pedestrian and **E** is the environment (spatial layout). This equation had shown that the pedestrian behavior is the expression of the pedestrian based on the surrounding environment(in this research context is the spatial layout) [13]. In this research, the pedestrian movement is proposed as the extended objective function to measure the fitness of the optimized GA based spatial layout design with full-fledged of operators for more optimization while providing the safest spatial layout for the pedestrian to move especially during panic situation.

2 Fitness Function: Autonomous Spatial Layout Design with Genetic Algorithm (GA)

Autonomous spatial layout design is the physical arrangement of a space based on the functional and architecture design paradigm. However, due to some limitation on the design's marker, the autonomous spatial layout design becomes non-fitness design and construct the less scalable and less functional floor plan [18]. Hence, the spatial layout design optimization was introduced for finding the feasible spatial layout based on the design's marker for the autonomous' construction. The most common spatial layout design optimization method is by using GA for designing an optimal spatial layout and will be used as a benchmark in floor plan design with scalable space usability [10, 19-21].

GA is the heuristic search with the natural evolution process adapted from genetic evolution in finding a better solution. This method is known as the optimization method. The original GA had highlighted the gene's transformation and modification as the process to acquire a new offspring with perfect and fittest value [18, 22]. Hence, this natural evolution process was adapted into autonomous spatial layout design process by optimizing the spatial layout arrangement and modification for finding the perfect fitness value via fitness function's quantitative measures. Fig. 2 shows the basic full-fledged of GA optimization algorithm and Fig. 3 shows the framework of spatial layout design optimization by using the full-fledged of GA operators from the previous research [10].

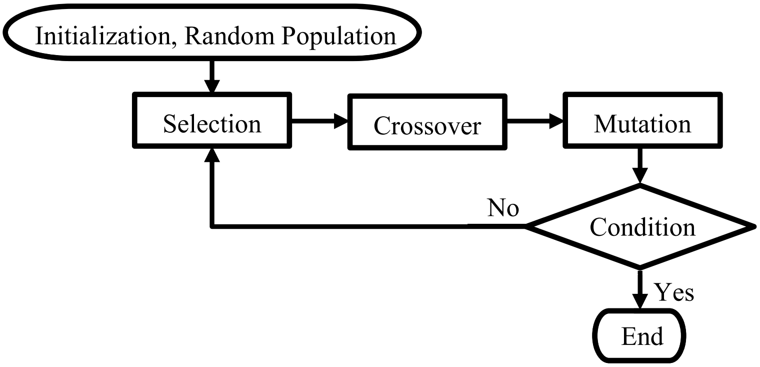


Fig. 2. Original framework of full-fledged GA

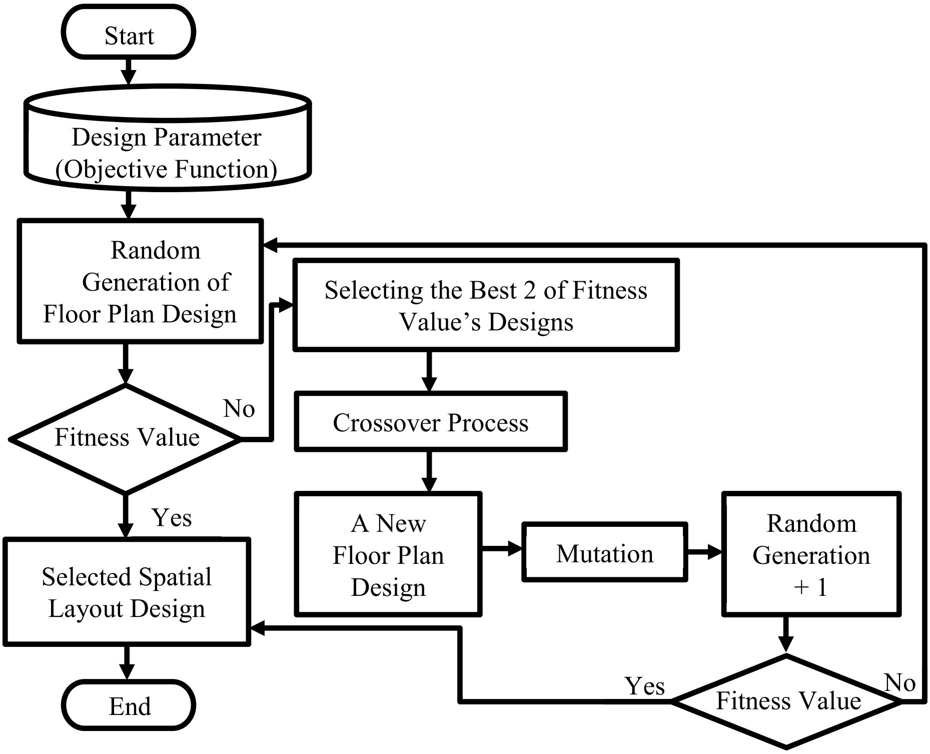

Start
Design Parameter (Objective Function)
Random Generation of Floor Plan Design
Selecting the Best 2 of Fitness Value's Designs
No
Fitness Value
Crossover Process
Yes
A New Floor Plan Design
Mutation
Random Generation + 1
Selected Spatial Layout Design
End
Yes
Fitness Value
No

Fig. 3. Framework for constructing a spatial layout design optimization with full-fledged of GA operators

Based on Fig. 3, the full-fledged GA from Fig. 2 was fused into the basic skeleton of building an autonomous spatial layout design. Based on Fig. 3, the objective function will be set up as a marker for the fitness function to distinctive the fittest value of the fitness values obtained [10]. From the random generation, two of the generated generation will be selected for the crossover operation for obtaining the fittest spatial layout design offspring. The generated offspring will be mutated by adjoining the offspring with the random new generation from the parent. The whole process will be done iteratively to find the highest fittest point of the fitness function for finding the most feasible spatial layout design that satisfy the objective function. These processes will be able to produce the fittest spatial layout design with the strong trait of the original plan layout [18, 23].

3 Pedestrian Movement Simulation: Microscopic Movement based CA Approach

During panic situation, pedestrian will have a great shock and will most likely to react randomly and abnormally [12, 16]. The pedestrian will try to get rid of any danger and physical collision by taking a counter measure in changing direction route's course to reach the safe place; self-organizing. This individual management promotes the microscopic movement based on the survival instinct. During survival rate prediction or pedestrian density observation, the pedestrian movement simulation able to play a great role on mimics the real pedestrian movement on a mock-up spatial layout design. Hence, through this research, one of the most established microscopic movement approach was used to mimics the self-organizing behavior of the pedestrian; Cellular Automata (CA) movement approach model.

CA is used to represent the pedestrian movement by using 3x3 homogenous grid cells. The grid cells can be occupied by any object such as people, obstacle, wall and incident spot within ***W***x***W*** grid layout. ***W*** is the height and width of the layout. There are some basic rules of CA;

a. For each time step, a cell can be occupied by only an object.
b. The pedestrian (object) must always move to towards ingress/ egress point.
c. The pedestrian and the moving incident spot must always move and gain one spot on cell for each time step.
d. The walls will remain permanent at the original spot while the incident spot must always spread based on the nature of the incident and pedestrian will always try to move towards the egress point while avoiding the physical collision between other pedestrian, walls, obstacles and incident spots.

The CA model approach is originally based on the von Neumann movement transition direction approach probabilities that consist of four types of movement; up, down, right, and left. Fig. 4(a) shows the basic movement transition for von Neumann movement approach. However, the movement transition approach was enhanced by using the Moore Neighborhood movement transition direction approach for more real-life simulation of the pedestrian movement. Fig. 4(b) shows the Moore Neighborhood movement transition based on 3x3 matric. These movement probabilities are based on the navigation direction as shown in Fig. 4(c).

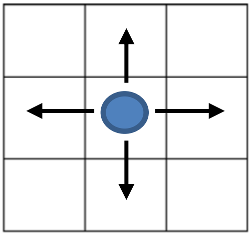


Fig. 4. (a) The von Neumann approach movement direction

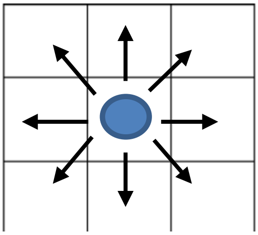


Fig. 4. (b) The enhanced Moore approach movement direction

(i-1, j-1)	(i, j-1)	(i+1, j-1)
(i-1, j)	(i, j)	(i+1, j)
(i-1, j+1)	(i, j+1)	(i+1, j+1)

Fig. 4. **(c)** The transition probabilities for enhanced Moore approach to represent the set of grid for movement simulation over a time step

4 Methodology: Pedestrian Simulation for the Fittest Spatial Layout Design

Spatial layout design plays a great role in structuring the pedestrian movement [13]. Hence, this research had used the previous research on autonomous spatial layout design with the GA optimization method [10] and had introduced the implementation of pedestrian movement simulation based on the cellular automata approach as the objective functions in determining the highest fitness value for designing the safe spatial layout. Fig. 5 shows the general framework of this research objective.

Based on Fig. 5, the basic GA method for spatial layout optimization was simulated using the main objective function; the non-overlapping of the obstacles' arrangement. The objective function was then extended at the final stage of the designing process by using the CA based pedestrian movement for optimizing the spatial layout design. The highest fitness value from the eight runs of the autonomous spatial layout design by using GA optimization method was collected and the selected designs were integrated with the CA based pedestrian movement simulation for finding the fitness value. The simulations were implemented in both normal and panic situation that been differentiate by using the movement speed. The pedestrian will move and change direction in speed 3 m/s for normal situation and 5 m/s for panic situation [24]. The fitness value of the simulation will determined the feasible spatial layout design that enable the pedestrian to move and escape especially during panic situation.

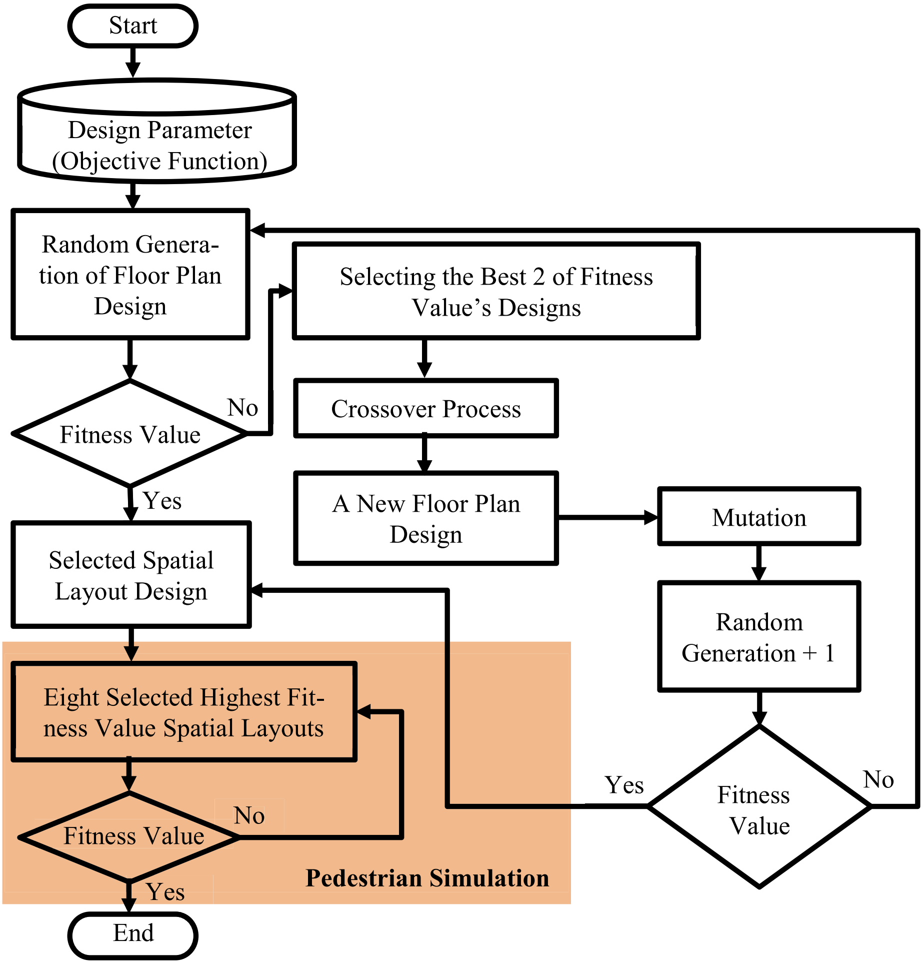

Start
Design Parameter (Objective Function)
Random Generation of Floor Plan Design
Selecting the Best 2 of Fitness Value's Designs
Crossover Process
No
Fitness Value
Yes
A New Floor Plan Design
Mutation
Selected Spatial Layout Design
Random Generation + 1
Eight Selected Highest Fitness Value Spatial Layouts
Yes
Fitness Value
No
No
Fitness Value
Yes
Pedestrian Simulation
End

Fig. 5. Autonomous spatial layout design optimization with GA operators and pedestrian movement simulation

5 Result

Experiments were conducted for this research to highlight the impact of the pedestrian movement as one of the objectives function for determining fitness value for designing the fittest spatial layout. Throughout this study, the experiment were conducted on the 14ft x 20ft spatial layout that was scaled into 14x20 grid cells and the environment were set into normal and panic situation for each run. The experiment will have eight of runs and gathered the best eight fittest value of optimized spatial layout design based on the non-overlapping of the obstacles as the objective function for GA operators approach. The eight fittest value from the optimized spatial layout using GA optimization method will be determined by running 500 iterations for each run with the cooperation of population, fitness function and GA operators; selection, crossover and mutation. Each of the fittest value will be run with 40 number of obstacles and number of pedestrian; 30, 60, 90, 120, and 150. Fig. 6 shows the basic floor plan to simulate the pedestrian movement on the selected space.

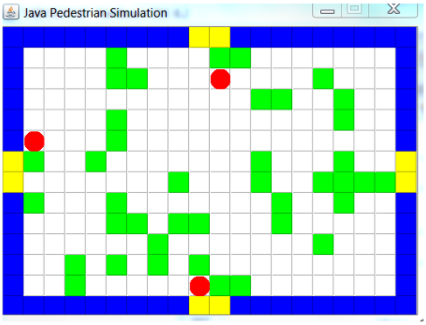


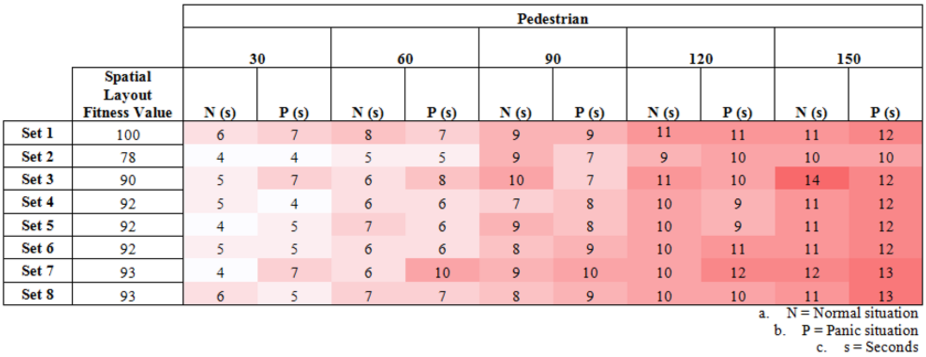
Fig. 6. The simulation of pedestrian movement in a spatial layout design

Note- *The blue color cells are the walls, the yellow color cells are the exit points, the green color cells are the static obstacles, the white color cells are the empty floor and the red color cells are the cells that occupied by the pedestrian.*

Based on Fig. 6, the pedestrian will be simulated for each run and the pedestrian will be randomly generated in the spatial layout for each run in every situations. Table 1 shows the result of pedestrian movement during evacuation in normal and panic situation in time (seconds) by using Moore Neighborhood movement transition direction as the direction approach.

Based on Table 1, the result had been interpreted to represent the heat map for showing the time (s) taken by the pedestrian to escape from the spatial layout. The heat map table shows that the fittest optimized spatial layout (Set 1) with 100 fitness value had shown the slower movement of the pedestrian to evacuate from the spatial layout, followed by the other highest 90% and above percentage of fitness value's sets; Set 3, Set 4, Set 5, Set 6, Set 7, and Set 8. However, the lowest GA optimized spatial layout design's fitness value (Set 2) with 78 fitness value had shown the fastest movement of the pedestrian compare to other fittest sets.

Table 1. Experiment result on the implementation of pedestrian movement simulation as the objective function



		Pedestrian									
		30		60		90		120		150	
	Spatial Layout Fitness Value	N (s)	P (s)	N (s)	P (s)	N (s)	P (s)	N (s)	P (s)	N (s)	P (s)
Set 1	100	6	7	8	7	9	9	11	11	11	12
Set 2	78	4	4	5	5	9	7	9	10	10	10
Set 3	90	5	7	6	8	10	7	11	10	14	12
Set 4	92	5	4	6	6	7	8	10	9	11	12
Set 5	92	4	5	7	6	9	8	10	9	11	12
Set 6	92	5	5	6	6	8	9	10	11	11	12
Set 7	93	4	7	6	10	9	10	10	12	12	13
Set 8	93	6	5	7	7	8	9	10	10	11	13

a. N = Normal situation
b. P = Panic situation
c. s = Seconds

Based on this heat map table, it is prove that the fitness function based on the spatial layout design optimization will able to design a good spatial layout with the structural and functionality values as the design's marker, in this research case study, the non-overlapping of the obstacles' arrangement. However, based on the result from the pedestrian movement simulation on Table 1, the optimized spatial layout design will also be able to affect the pedestrian movement. The highest fitness value of the GA optimized spatial layout also be able to cause the pedestrian to face a great challenge by changing route repeatedly due to high physical collision and facing the bottleneck effect especially near to the exit points for escaping from the spatial layout. Hence, as the analysis of the result in Table 1, the implementation on pedestrian movement as the objective function to find the fitness values can be a great value added for the autonomous spatial layout design to find the highest fitness value with high assurance of pedestrian's survival.

6 Conclusion

The Genetic Algorithm (GA) optimization method is able to optimize the spatial layout design via the full-fledged implementation of GA operators; selection, crossover and mutation for reaching the non-overlapping state of the obstacles' arrangement. However, the fitness value based of the physical arrangement on the spatial layout will not be able to ensure the safety of the pedestrian on the spatial layout. Hence, this research had proposed the usage of CA based pedestrian movement simulation as one of the objective function in finding the fitness value for a safe autonomous spatial layout design. This proposed method able to distinctive the high fitness values of the optimized GA based spatial layout design by simulating the CA based pedestrian movement for moving out from the spatial layout. The result shows that the lowest fitness value of GA based spatial layout design had enabled the pedestrian to escape faster than the highest fitness value of GA based spatial layout design.

Acknowledgement

Research experiment reported here is pursued under the Fundamental Research Grant Scheme (FRGS) by Ministry of Education Malaysia for "Enhancing Genetic Algorithm for Spatial Layout Design Optimization with Pedestrian Simulation in a Panic Situation" [203.PKOMP.6711534], Research University Grant (RUI) by Universiti Sains Malaysia for "Bio-Inspired Optimization Algorithm for Spatial Layout Design of Cul-de-sac Area using Microscopic Pedestrian Movement in a Panic Situation" [1001.PKOMP.8014073] and Bridging Grant by Universiti Sains Malaysia for "Pedestrian Simulation Model for Clogging Detection and Survival Prediction in a Fire Spreading Situation" [304.PKOMP.6316019]. The preliminary study of this research is supported under the Short Term Grant Scheme by Universiti Sains Malaysia for "Pedestrian Simulator and Heuristic Search Methods for Spatial Layout Design" [304.PKOMP.6313169].

References

1. Bhatt, M., et al. Artificial Intelligence for Predictive and Evidence Based Architecture Design. in AAAI. 2016.
2. Indraprastha, A. and M. Shinozaki, Computational Models for Measuring Spatial Quality of Interior Design in Virtual Environment. Building and Environment, 2012. **49**: p. 67-85.
3. Regateiro, F., J. Bento, and J. Dias, Floor Plan Design using Block Algebra and Constraint Satisfaction. Advanced Engineering Informatics, 2012. **26**(2): p. 361-382.
4. Mustafa, F.A. and A.S. Hassan, Mosque Layout Design: An Analytical Study of Mosque Layouts in the Early Ottoman Period. Frontiers of Architectural Research, 2013. **2**(4): p. 445-456.
5. Arnolds, I.V. and S. Nickel, Multi-Period Layout Planning for Hospital Wards. Socio-Economic Planning Sciences, 2013. **47**(3): p. 220-237.
6. IKEA. IKEA Home Planner Tool. 2016; Available from: http://www.ikea.com/ms/en_US/rooms_ideas/splashplanners_new.html.
7. Planner5D. Home Design and Interior Decor in 2D & 3D. 2017; Available from: https://planner5d.com/.
8. RoomSketcher. RoomSketcher Home Designer. 2017; Available from: http://www.roomsketcher.com/.
9. Zhu, Y., Y. Zhu, and H. Liao. A study of spatial layout design of fishing vessels. in Computer-Aided Industrial Design & Conceptual Design, 2009. CAID & CD 2009. IEEE 10th International Conference on. 2009.
10. Hassan, F.H., et al., Incorporating Genetic Algorithm Operators in Optimization Spatial Layout Design, in Proceedings of the International Conference on Advances in Image Processing. 2017, ACM: Bangkok, Thailand. p. 179-183.
11. Huixian, J. and Z. Shaoping. Navigation system design of fire disaster evacuation path in buildings based on mobile terminals. in 2016 11th International Conference on Computer Science & Education (ICCSE). 2016.
12. Sime, J.D., Crowd psychology and engineering. Safety Science, 1995. **21**(1): p. 1-14.
13. Kihlstrom, J.F., The person-situation interaction. The Oxford handbook of social cognition, 2013: p. 786-805.
14. Jay, B.N., Tahfiz did not have fire exit; bodies found piled on top of each other, in New Straits Times. 2017, New Straits Times Press (M) Berhad: Malaysia.

15. Four in a Family Killed in Fire, in The Star Online. 2017, Star Media Group Berhad (ROC 10894D).
16. Lu, X., et al., Impacts of Anxiety in Building Fire and Smoke Evacuation: Modeling and Validation. IEEE Robotics and Automation Letters, 2017. **2**(1): p. 255-260.
17. Yamin, M., H.M. Al-Ahmadi, and A.A. Muhammad. Integrating social media and mobile apps into Hajj management. in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). 2016.
18. Michalek, J. and P. Papalambros, Interactive Design Optimization of Architectural Layouts. Engineering Optimization, 2002. **34**(5): p. 485-501.
19. Hassan, F.H., S. Swift, and A. Tucker, Using Heuristic Search with Pedestrian Simulation Statistics to Find Feasible Spatial Layout Design Elements'. Journal of Algorithms, 2014. **2**(4): p. 86-104.
20. Hassan, F.H. and A. Tucker, Automatic Layout Design Solution, in Advances in Intelligent Data Analysis X: 10th International Symposium, IDA 2011, Porto, Portugal, October 29-31, 2011. Proceedings, J. Gama, E. Bradley, and J. Hollmén, Editors. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 198-209.
21. Hassan, F.H. and A. Tucker. Using Uniform Crossover to Refine Simulated Annealing Solutions for Automatic Design of Spatial Layouts. in IJCCI (ICEC). 2010.
22. Narahara, T. and K. Terzidis, Multiple-Constraint Genetic Algorithm in Housing Design. Synthetic Landscapes 25th ACADIA, 2006.
23. Renner, G. and A. Ekárt, Genetic Algorithms in Computer Aided Design. Computer-Aided Design, 2003. **35**(8): p. 709-726.
24. Helbing, D., I. Farkas, and T. Vicsek, Simulating dynamical features of escape panic. Nature, 2000. **407**(6803): p. 487-490.

QoE Enhancements for Video Traffic in Wireless Networks through Selective Packet Drops

Najwan Khambari[1] and Bogdan Ghita[2]

[1] Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
najwan@utem.edu.my
[2] Plymouth University, Plymouth, United Kingdom
bogdan.ghita@plymouth.ac.uk

Abstract. This paper proposes a queuing technique for important video frame packets with the objective to improve the performance of video transmission as perceived by the end users, across the IEEE 802.11e network. The proposed mechanism preserves the video Quality of Experience (QoE) by avoiding the I-Frames transmitted as part of the Group of Pictures (GoP) from being dropped during queue congestion. The method is evaluated using the NS-3 simulator with the Evalvid module and the results demonstrate the video flows will have better in Mean Opinion Score from the subjective evaluation point of view compared to the original IEEE 802.11e queueing.

Keywords: IEEE 802.11e; EDCA; QoE; NS-3; Evalvid.

1 Introduction

Wireless Local Area network (WLAN) has been in strong and growing demand since the early 1990s [1] [2] [3] and remains the preferred network access while being widely used [4]. This is due to the low deployment costs, flexibility, and ease of setup that provides an excellent means for generic networking. As the network becomes more popular, the network load has become a critical issue. The paradigm now has shifted where WLAN, which was originally designed and responsible to carry Best Effort (BE) services are now being used to carry heavy, time-sensitive traffic especially video.

Video services through the Internet has been in demand since the mid-2000s, and the trend has been growing significantly [5]. Simultaneously, the popularity of the Internet led to integrating video communications into the BE packet networks. This actually has been predicted since WLAN and Internet has become an integral part of everyday life. In 2010, [6] projected that 69% of the mobile traffic are accounted for video traffic. Meanwhile in 2011, [7] reported that half of the total Internet traffic are video and this pattern will continue to grow. These predictions were confirmed in 2014, with users were statistically being observed to spend more time watching video on a smartphone in 2014 compared to the previous year (2013) by 19.06% [8]. In 2015, a report was released by [9] and predicted increase in video traffic across the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_29

Internet from 64% in 2014 to 80% by 2019. An increase of traffic from wireless and mobile devices was also predicted where it will represent 66% of IP traffic by 2019 compared to 46% in 2014.

Transmitting video traffic over dynamic environment such as wireless networks remains a challenging issue in spite of the progress made through the IEEE 802.11e framework and the associated research. Since WLANs are now primarily the main method to stream multimedia traffic especially by end users, relying on QoS metrics is no longer sufficient.

In order to evaluate the quality of the multimedia file as being perceived by the users looking to parameters beyond QoS is now crucial. Thus, the term of Quality of Experience (QoE) has been introduced to expand the evaluation from the measurement of end-to-end performance at the services level to the impact these parameters have on the users' perception of the transmitted video. This new approach is required to define performance measurement while considering users' subjective nature [10].

2 Related Works

To understand the current state of the art in wireless video QoE, this section will provide an overview of the video encoding process, the wireless queuing, and the efforts made to combine the two concepts in a communication architecture.

2.1 GoP in MPEG-4

MPEG-4 was developed to store and deliver multimedia content over the Internet [11]. MPEG-4 streams consist of three types of frames, namely I, P, and B, transmitted in a structure called Group of Pictures (GoP).

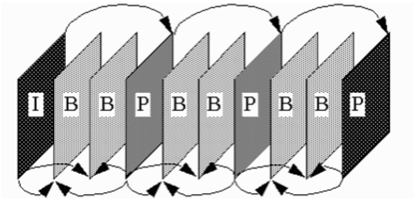


Fig. 1. GoP

The I-Frames are independent of the rest of the stream, as they do not depend on other frames to allow decoding of the video stream. Meanwhile P-Frames and B-Frames only contain video information updates compared to the I-Frames. Because of their content, P-Frames and B-Frames have smaller sizes in comparison with I-Frames; they are also depending on the content of the I-Frame in order to allow the encoding of the video at the receiver. I-Frame is the most important frame where P- and B-Frames within the same GoP cannot be decoded without the I-Frame.

2.2 IEEE 802.11e EDCA

In recent years, a lot of effort had been made to enhance the IEEE 802.11 network performance. In spite of its benefits, the original IEEE 802.11 standard specification did include a number of inherent challenges including supporting QoS for video traffic. To address this issue, IEEE amended the original IEEE 802.11 and formed Task Group E to enhance the MAC layer of IEEE 802.11 to support QoS. Hence a new amendment known as the IEEE 802.11e was introduced.

IEEE 802.11e introduced two new channel coordination functions: HCF Controlled Channel Access (HCCA) and Enhanced Distributed Channel Access (EDCA). HCCA is based on polling where the Access Point (AP) polls each Mobile Station (MS) to check whether it wants to send any data. In contrast, EDCA is a contention based channel that requires each MS to compete to gain access of the wireless channel.

EDCA is an upgraded version of the legacy Distributed Coordination Function (DCF) of IEEE 802.11 which incorporate several enhancements. Four different Access Categories (AC) were introduced: AC0, AC1, AC2, and AC3, where they were intended to refine the granularity of the queue, based on the traffic priority. The traffic priorities are grouped according to the type of services it carries. AC3 is solely used to stream voice (AC_VO), AC2 to stream video (AC_VI), best effort traffic in AC1 (AC_BE) and AC0 for background traffic (AC_BK). The priority ranks from the highest to the lowest, respectively. These streams will be treated differently in the EDCA mechanism. In this paper, focus was given on this particular feature where we further refine the AC_VI queue to manage the video traffic better.

The introduction of IEEE 802.11e EDCA has significantly improved the QoS provision for wireless network [12] [13] where it has been biasedly designed to be selective towards the high priority AC. However, several issues still need to be addressed before EDCA can really support QoS as well as QoE for video traffic. Multiple ACs in the queue causes internal collision where individual ACs competes. Moreover, the current standard treats all video the same regardless of the content type while prioritization on the different types of video frames are totally neglected.

2.3 Previous works

Over the years, significant amount of work has been done to improve the performance of video traffic in the IEEE 802.11e EDCA. To gauge the effectiveness, both objective and subjective video quality evaluation methods have been used. Peak Signal to Noise Ratio (PSNR) is an objective method that uses Mean Square Error (MSE) to compare the original and reconstructed video on an image-to-image comparison basis while Structural Similarity Index (SSIM) compares the video pixel-to-pixel. Some works has also involved subjective evaluation such as using Mean Opinion Score (MOS) to rate the original and reconstructed video.

Several papers improve the video transmission performance by adjusting the parameters of the IEEE 802.11e EDCA mechanism. These include the Transmission Opportunity (TXOP), Arbitrary Inter-frame Space (AIFS) and the Contention Window (CW). These techniques are quite popular as it is easier to implement while reduce the waiting time to access the wireless media for the high priority traffic.

However, they do not exploit the significance of specific traffic type such as video that needs to be dynamic due to the nature of its variability in data rates. For example, [14] aimed to enhance the QoS level in EDCA for medical video communication. This is to provide the required medical-grade QoS of various medical applications to ensure accurate visualization of the patient condition. AIFS were adjusted to improve network performance to achieve the desired packet delay and the ratio of late/ receive packets.

Meanwhile, several papers considered a cross-layer approach. Packets are individually tagged according to their priority level. Through the proposed technique, queueing systems may identify the types of the video packets and treat them selectively. For example, [15] proposed a static mapping where I-Frame is mapped to AC_VI while P- and B-Frames are mapped to AC_BE and AC_BK respectively. Through this technique, I-Frames are given the highest priority and this prevents them from being dropped whenever the queue is congested. However, when the video traffic load is light, channeling P- and B-Frames to AC_BE and AC_BK causes unnecessary delay.

An alternative approach was used in [16] by introducing dynamic video frame mapping. The mapping is based on the significance of the video data and the traffic load. In video traffic, I-Frames can cause a sudden burst due to the size of the I-Frame itself. This will lead to a sudden congestion in the queue. Therefore, packets that are incoming toward the queue will he highly likely to be dropped. Through this proposed adaptive mapping, any incoming video packet will be channeled to AC_BE or AC_BK if AC_VI is full to avoid it being dropped. The system however, will not interfere with the AC_VO, the highest priority AC. Through the evaluation test, the video traffic PSNR of the proposed mechanism can be seen to increase by 8.11% to 11.80%.

Also in the context of cross layer approaches, [17] introduced a design to ensure better QoE for H.264 with Scalable Video Coding (SVC) video. The proposed mechanism involves three layers - APP, MAC and PHY. The APP and PHY layers are made aware of each other's condition. The receiver will send ACK to indicate whether packets were received correctly. Based on the ACK records, online QoS-QoE mapping was proposed. The PHY layer adapts to provide unequal error protection for each video layer based on the proposed mapping. Meanwhile, the APP layer is updated on the buffer starvation of the channel by the PHY and adjusts its rates accordingly. The simulation scenarios indicate that the architecture successfully avoids buffer starvation while handling channel and buffer fluctuations to achieve a 30% increase in video capacity. However in this experiment, only SVC videos were considered and the exact wireless network technology was not mentioned.

Adaptive Mapping Mechanism (AMM) was introduced in [18], based on the earlier work from [16]. Through a number of enhancements, AMM is capable to check the congestion level of all the voice, video, best effort and background queues before assigning video packets to the other queues. The motive of the enhancement is to stop the video traffic from monopolizing the access control for the station. The study also introduces differentiated queuing depending on frame type, with I-Frames assigned to the highest AC priority, AC_VO and, depending on the mapping control module, P-

Frames assigned to AC_VO or AC_VI and B-Frames assigned to AC_VI or AC_BE. Through simulation, the AMM proved to have less packet loss ratio.

Besides that, recent studies also suggested on developing the ability of IEEE 802.11e to be able to offload time sensitive packets on other networks such as cellular [19] and IEEE 802.16 WiMax [20]. Though it may improve the QoE of certain services, it can only be implemented in an ultra-dense deployment and that, it is dependent on the performance of the network it offloads to.

Most of the previous study suggested that I-Frames should be protected from being dropped in a video transmission to achieve better QoE. Although I-Frame's priority can be increased by channeling it to AC_VO, P- and B-Frames still have the old priority. This means I-Frame arriving early at the receiver does not give any effect as the video still cannot be reconstructed until the consecutive P and B-Frames arrive as well. Moreover, depending on scenario, this will probably cause delay to video traffic especially in a scenario where voice traffic is high and thus P and B-Frames needs to give way to the voice traffic. Theoretically a video can be reconstructed without P and B-Frames. However, if that is the case in this scenario, bandwidth has been wasted to transport the unused P and B-Frames to the receiver.

3 Proposed Method

In Section 2.1, it has been discussed that the I-Frame of a GoP is the most important frame in an MPEG-4 video, therefore protecting the I-Frame ensures a high probability of the video quality being preserved. In this paper, focus is given in a scenario where the AC_VI queue is congested. This can occur especially when there is a sudden increase in video flow or when the network traffic is highly loaded.

In a queue, packets can be dropped due to several reasons. Two of the reasons are discussed here. The first reason is due to that the queue is full. In this situation, incoming packets will be dropped as there are no spaces to allocate new packets in the queue. Secondly is due to the expiry time (timeout) of a packet. In this scenario, packets that are already in queue for a certain period of time are to be dropped due to the maximum duration a packet can exist in the queue. It is also known as the maximum delay (MaxDelay) a packet can wait in a queue before being transmitted. If the maximum time of the packet is reached (MaxDelay reaches zero) and the packet is still in the queue, it will be dropped from the queue. This is to avoid the 'stale' packet to hinder the transmission of the other packets behind it. This is a normal practice in IEEE 802.11 and is reflected in the simulation tool that we are using in this study.

Given the above scenarios, the objective of the proposed scheme is to avoid I-Frames from being dropped by discarding the B-Frames from the AC_VI queue during congestion. In a separate work conducted [21], we have discovered that B-Frame packets can be dropped to a certain level without notably affecting the PSNR.

In the proposed mechanism, I-Frame Manager (IFM), packets with P- and B-Frame information (PktP and PktB) are queued using the default EDCA queueing mechanism, DefaultQueue (DQ), but packets carrying I-Frame information (PktI), are treated differently by going through a three-step queue testing. First is the AC_VI congestion level. If the AC_VI queue is not full, PktI will be queued as usual using DQ. However if the AC_VI queue is full, the second queue testing will take place, which is to check whether there is a PktB of that particular video flow available in the AC_VI to be removed. If it is positive, the PktB will be removed from the queue and will be replaced by the incoming PktI. The third step involves checking the MaxDelay of PktI in the queue. If there is any PktI that has been identified to be near to its MaxDelay, any PktB in front of the PktI will be removed so that the PktI will not be dropped due to timeout. The mechanism flow chart can be shown as in Fig. 2.

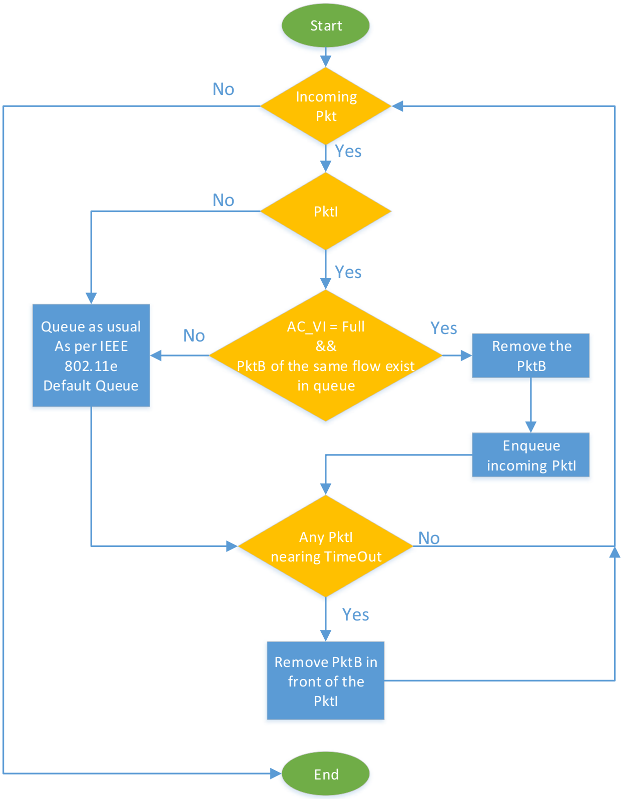

Start
Incoming Pkt
No
Yes
PktI
No
Yes
Queue as usual As per IEEE 802.11e Default Queue
No
AC_VI = Full && PktB of the same flow exist in queue
Yes
Remove the PktB
Enqueue incoming PktI
Any PktI nearing TimeOut
No
Yes
Remove PktB in front of the PktI
End

Fig. 2. Flowchart of the proposed IFM

4 Network Simulation Setup

To assess the IFM mechanism against DQ, a simulation study was conducted based using NS-3.22. Four MS and an AP were involved in the infrastructure mode of the wireless network scenario. During the course of transmission, only one sender and one receiver is involved. Since every station including the AP is within each other's range, there are no hidden terminals involved. QoS is enabled in all of the stations involved to simulate the IEEE 802.11e medium access control and the dedicated traffic-typed based queueing system. Meanwhile, to minimize the control packet in the wireless environment, Request to Send/ Clear to Send (RTS/ CTS) is disabled.

The experiment uses the "highway.yuv" video sequence, which can be obtained publicly [22]. Since the proposed mechanism is intended to address the issue of video queue congestion, this video was selected as it is made up of the highest number of frames from the available public video samples. The video therefore, will likely to cause queue congestion. The video is in a CIF format with a resolution of 352 x 288.

Evalvid [23] was imported into the NS-3 simulator to simulate real video frames transmission across the simulation scenario. Several other tools were also used for the simulation such as ETMP4, FFMPEG and PSNR to encode/ decode and evaluate the video quality used in the simulation. The video used in this experiment is encoded with a GoP size of 9.

5 Results and Findings

Following the simulations, the video streams collected at the receiver are evaluated in terms of visual comparison and subjective evaluation between IFM and the default queueing mechanism of the IEEE 802.11e. As part of the subjective evaluation, Mean Opinion Score (MOS) from a group of subjects were collected where they were required to watch the video as being perceived by the receiver's end.

5.1 Subjective Evaluation

There are many tools to evaluate the quality of a video objectively such as PSNR and Structural Similarity (SSIM). However, the most accurate and proven technique is to evaluate a video through subjective analysis [24]. A group consisting of 100 subjects was requested to watch the video that has been reconstructed as being perceived by the receiver. The video are evaluated by the subjects through MOS, based on the Absolute Category Rating (ACR) in line with the ITU-T guidelines and recommendations which can be referred in [25]. The MOS scale is shown as in Table 1 below.

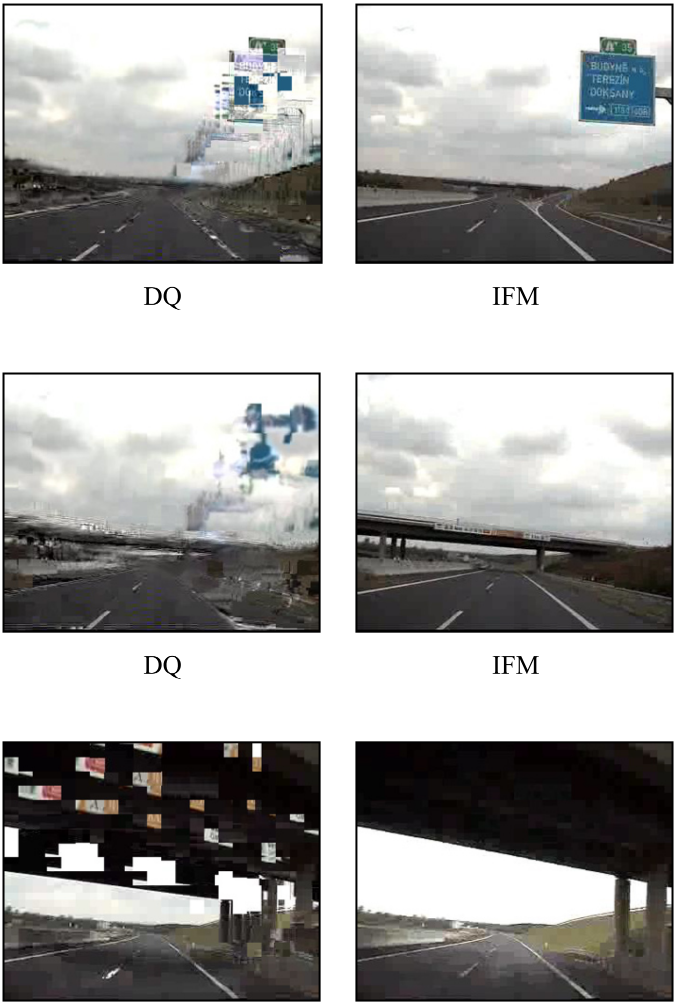
Table 1. Mean Opinion Score (MOS) scale of a video quality evaluation

Mean Opinion Score	
5	Excellent
4	Good
3	Fair
2	Poor
1	Unacceptable

For both original and reconstructed video that implements IFM, an average value has been taken from the scores that were provided by the subjects. Therefore, from the results of the MOS, the video which was using the DQ scored **2.51** while the proposed method scored at **3.84**. In terms of interpretation of the video quality mapped from the MOS scale, the video showed improvements from fairly "Poor" to nearly "Good" which can be seen from the visual comparison in the next section.

5.2 Visual Comparison

Fig. 3 presents snapshots of video quality between three different frames, that uses the DefaultQueue (DQ) and the proposed mechanism (IFM).



DQ IFM

DQ IFM

DQ IFM

Fig. 3. Visual comparison of the reconstructed video at the receiver's end

The quality of video delivered using IFM significantly improved the video quality in comparison to DQ.

5.3 Conclusion and Future Work

This paper proposes a selective queuing method that enhances the performance of MPEG-4 video transmission in the IEEE 802.11e environment. The mechanism proposed preemptively drops B-Frame packets in the queue on two occasions; during congestion and I-Frame packets nearing timeout in a queue to prioritize the I-Frame packets. The validation experiments demonstrated that the proposed mechanism has the ability to provide enhancements in video transmission and thus offers better video QoE in the IEEE 802.11e.

This is a work in progress where in the future development, the proposed system would be able to differentiate the content of the video in a multiple video flow environment and currently being used as the metric to compare the effectiveness of the proposed scheme. Here, subjective evaluations were used to provide an effective method to reflect how the proposed technique has improved the video quality. In the future, objective metrics such as PSNR and SSIM will be taken into account to provide a more holistic result as to the effectiveness of the proposed scheme besides testing on other video formats such as Scalable Video Coding (SVC).

Acknowledgement

This project is part of a PhD research currently being carried out at Centre for Security, Communications and Network Research (CSCAN), Plymouth University, U.K. The deepest gratitude and thanks to Universiti Teknikal Malaysia Melaka (UTeM) and the Malaysian Ministry of Higher Education for funding this PhD research.

References

1. Braden R, Clark D, Shenker S.: RFC1633: Integrated Services in the Internet Architecture: an Overview. IETF RFC 1633, July 1–28 (1994)
2. Khan A, Sun L, Ifeachor E: Content Clustering Based Video Quality Prediction Model for MPEG4 Video Streaming over Wireless Networks. Journal of Multimedia 4:228–239 (2009).
3. Arabi M, Ghita B, Wang X (2010) Improving Fairness in Ad Hoc Networks through Collision Rate Control. In: International Network Conference. pp 51–59, Heidelberg, Germany (2010).
4. Kosek-Szott K, Natkaniec M, Szott S: What's new for QoS in IEEE 802.11? IEEE Network (6), 95-104 (2013).
5. Vassiliou V, Antoniou P, Giannakou I, Pitsillides A.: Requirements for the Transmission of Streaming Video in Mobile Wireless Networks. In: International Conference on Artificial Neural Network. pp 528–537, Athens, Greece (2006).
6. Stanley M Internet Trends (2010).

7. Cisco: Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update, 2010 – 2015 (2011).
8. Nielsen: The Total Audience Report (2014).
9. Cisco: Cisco Visual Networking Index : Forecast and Methodology , 2014 – 2019 (2015)
10. Jain RMedia vision - Experience Isn't Only Screen Deep. IEEE Multimedia 10, (1),80–81 (2003).
11. Lingfen S, Is-Haka M, Emmanuel J, Emmanuel I.: Guide to Voice and Video over IP. Springer (2013).
12. Anitha A, Jayakumari J.: Performance Analysis of WLAN Under Variable Number of Nodes Using the Adjustable Parameters in EDCA. Journal of Theoretical and Applied Information Technoly 60(2), 351–357(2014).
13. Abu-khadrah A, Zakaria Z, Othman M.: Evaluate QOS parameters for VOIP using IEEE 802.11 (DCF) and IEEE 802.11e (EDCA). Australian Journal of Basic and Applied Science 8, 265–272 (2014).
14. Son S, Park K, Park E: Adaptive tuning of IEEE 802.11e EDCA for medical-grade QoS. In: 2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, pp 650–651, Da Nang, Vietnam (2013)
15. Ksentini A, Naimi M, Gueroui A.: Toward an Improvement of H.264 Video Transmission over IEEE 802.11e Through a Cross-Layer Architecture. IEEE Communications Magazine 44(1), 107–114 (2006).
16. Lin C-H, Shieh C-K, Ke C-H.: An Adaptive Cross-Layer Mapping Algorithm for MPEG-4 Video Transmission over IEEE 802.11e WLAN. Telecommunication Sysemt 42, 223–234 (2009).
17. Abdel Khalek A, Caramanis C, Heath RW A Cross-Layer Design for Perceptual Optimization Of H.264/SVC with Unequal Error Protection. IEEE Journal on Selected Areas in Communication 30, 1157–1171 (2012).
18. Yao XX-W, Wang W, Yang S.: IPB-frame Adaptive Mapping Mechanism for Video Transmission over IEEE 802.11e WLANs. ACM SIGCOMM Computer Communication Reviews 44, 5–12 (2014).
19. Bouali F, Moessner K, Fitch M.: A Context-Aware User-Driven Strategy to Exploit Offloading and Sharing in Ultra-Dense Deployments. IEEE International Conference on Communications, pp 1-7, Paris, France (2017).
20. Ahmed N-ER: A QoS Based Algorithm for the Vertical Handover between WLAN IEEE 802.11e and WiMAX IEEE 802.16e. International Journal of Computing and Digital System 7(1) 11-22, (2017).
21. Khambari N, Ghita B, Sun L.: QoE-Driven Video Enhancements in Wireless Networks Through Predictive Packet Drops. In: International Conference of Wireless Mobile and Computer Network Communications, pp 355-361, Rome, Italy (2017).
22. YUV QCIF reference videos (lossless H.264 encoded). http://www2.tkn.tu-berlin.de/research/evalvid/qcif.html. Last accessed 2016/2/15
23. Klaue J, Rathke B, Wolisz A.: EvalVid - A Framework for Video Transmission and Quality Evaluation. In: International Conference on Modelling Techniques and Tools for Computer Performance Evaluation. pp 255–272, Berlin, Germany (2003).
24. Reckwerdt B, Clarity V White Paper : Understanding MOS , JND , and PSNR
25. ITU-T: Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment(2014).

Preventing Denial of Service Attacks on Address Resolution in IPv6 Link-local Network: AR-match Security Technique

Ahmed K. Al-Ani[1], Mohammed Anbar[1], Selvakumar Manickam [1], Ayman Al-Ani[1] and Yu-Beng Leau[2]

[1] National Advance IPv6 Center, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia
[2] Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
ahmedkhallel91@nav6.usm.my

Abstract. Address resolution (AR) process, one of the important neighbor discovery protocol (NDP) functions, aims to obtain the corresponding relationship between Internet protocol and media access control addresses. This process uses two NDP messages, neighbor solicitation (NS) and neighbor advertisement (NA) messages, which are unsecure by design. In addition, the target address is revealed in the traditional AR process. Thus, any malicious node on the same link can modify the message and launch denial of service (DoS) attacks. The current mechanisms suffer from high-complexity issue or other forms of security issues that can induce DoS attack on AR in IPv6 link-local network. To overcome these limitations, this work proposes AR-match technique to secure AR process by hiding the target address by using a hash function algorithm and adding a new option named AR-match, which is attached to each NS and NA message for them to become NS- and NA-match messages, respectively. AR-match technique can provide a high security with less complexity and will completely prevent DoS attacks during AR in the IPv6 link-local network.

Keywords: Address Resolution, Neighbor Discovery Protocol, IPv6 Link-local Network, Network Security.

1 Introduction

Internet Protocol version six (IPv6) was introduced to replace the actual Internet Protocol version Four (IPv4) protocol. IPv6 brings many new capabilities and improvements considering simplicity, routing speed, quality of service and security [1]. In comparison to IPv4, IPv6 improves mechanisms for assuring a secure and confidential transfer of information. IPv6 comes with a new protocol to manage the communication among nodes that locate on the same link, this protocol called Neighbor Discovery Protocol (NDP). NDP consists of a set of messages and procedures that determine the relationships between nodes (routers and hosts) and their neighbor's in the same network [2]. NDP substitutes several protocols used in IPv4 such as Router Discovery, Address Resolution (AR), Internet Control Message Protocol (ICMP) and ICMP Redirect. NDP also provides new functions such as Duplicate Address Detec-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_30

tion (DAD) and Neighbor Unreachability Detection (NUD) [3]. Similarly, IPv6 NDP allows nodes to identify their neighbor s on the same LAN and publicise their existence to other neighbor's [4]. One of the interesting points in IPv6 networks is the capability of all nodes to automatically configure their addresses using stateless address auto-configuration (SLAAC) [5] without deploying any state-full configuration protocol, i.e. the Dynamic Host Configuration Protocol for IPv6 (DHCPv6). SLAAC provides the power to an IPv6 host to generate link-local and global addresses without manual intervention [6]. NDP accomplishes its functions by using five ICMPv6 messages as shown in Table 1:

Table 1. NDP Messages

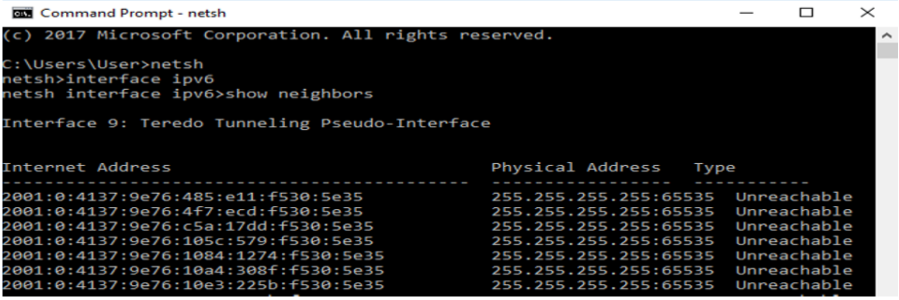
Message Type	Transmission		Purpose
	From	To	
Router Solicitation (RS) – 134	Node	Router	Requests router to send RA message
Router Advertisement (RA) – 134	Router	Node	Transmits periodically or in response to RS to inform nodes about network configuration information
Neighbor Solicitation (NS) – 135	Node	Node	Request another node its media access control address (MAC) address and other functions such as NUD and DAD.
eighbour Advertisement (RA) – 136	Node	Node	Transmits in response to NS message or to advertise the new address.
Redirect Message (RM) – 137	Router	Node	Used to redirect the node's traffic from path to another better path.

Despite these improvements, network security remains a very important issue because security threats and attack types can affect IPv6 network. This paper proposes a new technique, namely, AR-match to secure one of the NDP functions in IPv6 link-local network. The rest of this paper organised as follows. Section 2 explains the main process of AR. AR security issues are elaborated in Section 3. Section 4 discusses the limitations of current existing mechanisms. Section 5 discusses the proposed technique. Sections 6 and 7 discuss the expected results, future work and conclusion, respectively.

2 Address Resolution (AR) Process

One of the major functions of NDP is the Address Reclusion process, AR in short. When a node has a unicast packet to send to a neighbor, but does not know the neighbor's link-layer address, it performs address resolution [7]. ARP executes by using two of NDP messages i.e. Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages. Each host in the IPv6 link-local network has a neighbor cache table containing all the hosts neighbors locate on the same link [3]. Neighbor cache table can shows by using these commands in the Command Prompt in Windows operating system, as shown in figure 1:

netsh → interface ipv6 → show neighbors



```
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\User>netsh
netsh>interface ipv6
netsh interface ipv6>show neighbors

Interface 9: Teredo Tunneling Pseudo-Interface


Internet Address                              Physical Address   Type
--------------------------------------------  -----------------  -----------
2001:0:4137:9e76:485:e11:f530:5e35            255.255.255.255:65535  Unreachable
2001:0:4137:9e76:4f7:ecd:f530:5e35            255.255.255.255:65535  Unreachable
2001:0:4137:9e76:c5a:17dd:f530:5e35           255.255.255.255:65535  Unreachable
2001:0:4137:9e76:105c:579:f530:5e35           255.255.255.255:65535  Unreachable
2001:0:4137:9e76:1084:1274:f530:5e35          255.255.255.255:65535  Unreachable
2001:0:4137:9e76:10a4:308f:f530:5e35          255.255.255.255:65535  Unreachable
2001:0:4137:9e76:10e3:225b:f530:5e35          255.255.255.255:65535  Unreachable
```

Fig. 1. IPv6 Neighbor Cache Table

When host A wants to communicate or send a data packet to host B, host A should firstly check whether the MAC of B is available in the cache; otherwise, host A should perform AR process. To do so, host A needs to multicast NS message to solicited-node multicast address (SNMA) based on the last 24 bits of host B IP address asking host B to reply with its MAC address. All hosts in the local area network receive the NS message. However, only host B provides an AR reply through NA message which contains its IP and MAC addresses. After receiving a reply, host A updates its cache and sends a data packet to host B according to the MAC address in the cache [8]. Fig. 2 shows the AR process between hosts A and B.

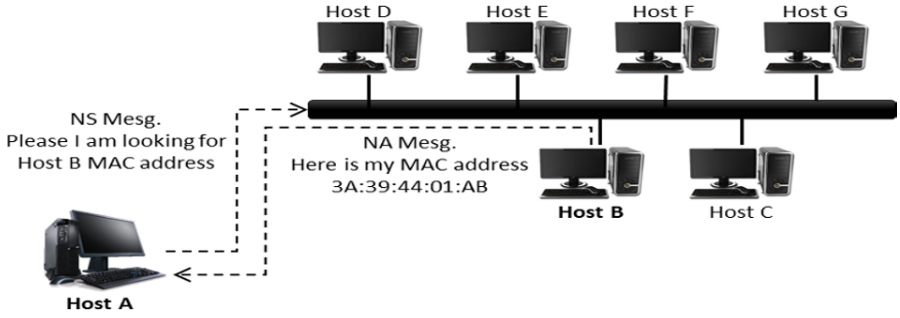


Fig. 2. AR Process

3 Security Issues of AR Process in IPv6 Link-local Network

Although AR is simple and efficient, security risks remain. Firstly, all hosts on the same IPv6 link can join the AR process. Therefore, any malicious node on the link can disrupt the AR process by sending a fake NA message because NDP messages are unsecure by design [4], [9]–[11]. Secondly, AR process lacks a verification mechanism [12]. For instance, when host A needs to communicate with host B, the former does not check whether the response NA message is valid or fake. Host A updates its neighbor cache table once NA messages are received with an invalid MAC address. Therefore, this process provides convenience for DoS attack. The process of a common attack is as follows:

When host A wants to communicate with host B, host A must perform AR process by sending NS message asking host B for its MAC address. Host C may send NA message as a response to NS message, pretending to be host B. After receiving the NA message, host A does not know that it is a fake message and mistakenly updates its neighbor cache table, regarding host C as host B, and sends to host C the data packet that should have been sent to host B. Fig. 3 shows the ARP spoofing attack.

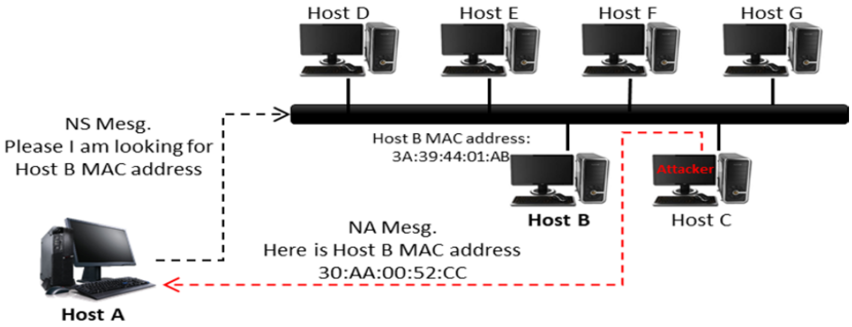


Fig. 3. AR Spoofing Attack

The main problem of this research is summarised as follows. (i) Revealing target address in NS message allows the attacker to launch DoS attack on the target host. (ii) Two NDP messages, namely, NS and NA messages, which are unsecure by design, are used during AR process in IPv6 link-local network. Thus, any attacker can modify the NA message and send it as a response to the NS message. (iii) The current existing mechanisms such as SeND, SAVA, IDS-mechanisms and SAVI suffer from either high-complexity issue or other form of security issues that can induce DoS attack on AR process in IPv6 link-local network.

4 Related Work

There are some researches attempt to secure AR process in IPv6 link-local network by using third party as an agent to monitor all the network behaviour and if there is any abnormal behaviour will send an alarm or report to administrator. For example, researches [13]–[15] used Intrusion Detection System (IDS) to secure NS and NA messages during AR process. These researches introduce a method to monitor and record all <IP, MAC> mappings of the all hosts on the network for a long period. When the <IP, MAC> mapping in the address resolution message that the host sends out is not consistent with the record, a spoofing attack exist, thus, IDS will send an alarm to inform the administrator there is a spoofing attack and the administrator will deal with that attack. However, this method this method belongs to passive defense and increases network complexity and maintenance cost as cited in [10], [16].

Although RFC 4861 [17],[18] suggests using IP security (IPsec) to protect the NDP messages include NS and NA messages, the use of IPSec in neighbor discovery still faces several problems. The Internet Engineering Task Force (IETF) proposed SEcure- Neighbor Discovery (SEND) to enhance the security of NDP. SEND uses cryptographically generated address (CGA), digital signature, and time stamp to protect the NDP message [19]. However, key management remains a problem in encryption communications; the coexistence of SEND and NDP may also cause routing problems. Given its high complexity [20], [21], SEND remains to be in the experimental stage [22]. A logical problem in encrypted communication is that both parties must know the MAC address of each other before key change; acquiring the MAC address is the purpose of address resolution. If fraud occurs in the MAC-obtaining process, then the follow-up communication security is lost. Source Address Validation Architecture (SAVA) [23] is proposed to filter packets based on their source addresses. This method prevents attacks directly from the source and provides convenience for source address tracking, traceability, and network diagnosis and management. Source address validation implementation (SAVI) [24] is a protocol and a security mechanism that needs the entire network support, and remains in its experimental stage. Source address validation in autonomous regions needs a router support, which must complete several centralized computing to affect network performance. Also, SAVI is difficult to deploy because equipment manufacturers implement the Simple Network Management Protocol (SNMP) in various ways [25]–[27]

Most studies lack theoretical support. Single point of failure, hardware cost, operation complexity and other factors also restrict the development of AR process because they can induce DoS attack. This study proposes a new security technique named AR-match to resolve the address resolution security issue by introducing a new option called AR*match* to be attached to each NS and NA message for them to become NS- and NA-match message, respectively. The next section will explain in detail the proposed AR-match security technique.

5 Proposed AR-match Technique

5.1 Main Stages of AR-match Technique

The proposed technique aims to secure AR process in IPv6 link-local network by hiding the target address through redesigning NS and NA messages. This section describes the three main stages of the DAD-match technique, namely, secure target address stage, secure NS and NA message stage and DoS-AR prevention stage.

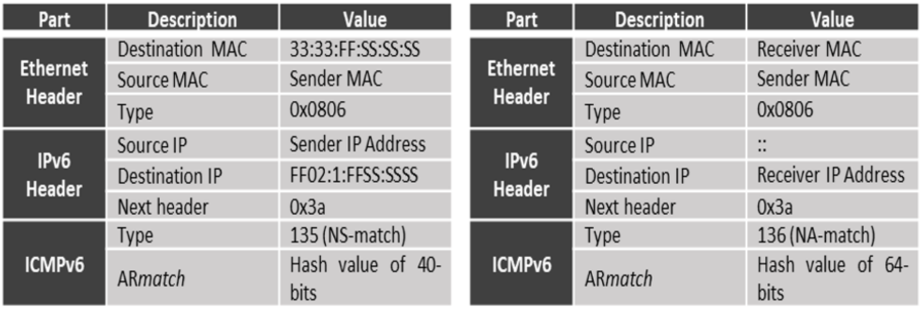
Secure Target Address (Stage One)

In this stage, the target address should be hidden securely by hashing the first 40 and 64 bits of its host portion. The hash value of the 64 bits is saved in the cache registry for later NA message verification. The hash value of the 40 bits is carried by the new option named AR*match* which will be attached to each NS and NA message for them to become NS- and NA-match message, respectively.

.

Secure NS and NA Messages Stage (Stage Two)

The AR process relies on two NDP messages (NS and NA messages) as mentioned previously. Therefore, NS and NA messages should be redesigned to secure the AR process because they are unsecure by design and cannot distinguish between valid and invalid messages. The AR-match technique introduces an option, named AR*match* option, which holds the hash value of the first 40 and 64 bits of host portion of target address for NS and NA verification purposes, respectively. The AR*match* option should be attached to each NS and NA for them to become NS- and NA-match messages, with type field values of 135 and 136, respectively. The NS-match message will have the 40-bit hash value which is going to be sent by sender. However, NA-match message will have the 64-bit hash value which is going to be sent by the receiver. Each message without the AR*match* option should be discarded. The verification of the message to differentiate whether it comes from a legitimate or illegitimate node is based on matching the hashing of the incoming hashing IP address with its self-hashing IP address. Fig. 4 shows the message format for NS- and NA-match messages.



Part	Description	Value
Ethernet Header	Destination MAC	33:33:FF:SS:SS:SS
	Source MAC	Sender MAC
	Type	0x0806
IPv6 Header	Source IP	Sender IP Address
	Destination IP	FF02:1:FFSS:SSSS
	Next header	0x3a
ICMPv6	Type	135 (NS-match)
	AR*match*	Hash value of 40-bits

Part	Description	Value
Ethernet Header	Destination MAC	Receiver MAC
	Source MAC	Sender MAC
	Type	0x0806
IPv6 Header	Source IP	::
	Destination IP	Receiver IP Address
	Next header	0x3a
ICMPv6	Type	136 (NA-match)
	AR*match*	Hash value of 64-bits

Fig. 4. AR-match Messages (NS-match and NA-match Messages) Format

DoS-AR Prevention Stage (Stage Three)

After the target host attaches the option AR*match* to the NS message to become an NS-match message which holds the hash value (Fig. 5), the target host should send to the SNMA address (FF02::1:FFSS:SSSS), in which S represents the last 24 bits of target IP address. All *existing* hosts on the same link receive the NS-match message, and the existing host should match the hash value. After computation, if the hash value matches, the AR process is performed, and the host can reply via NA-match message after hash the 64-bit host portion of target IP address and insert it to the AR*match* option, which should also be appended to the NA-match message. Upon receiving the NA-match message, the new host matches the 64-bit hash value. If matches are found, then the message comes from a legitimate node, and the target host can update its neighbor cache table and start to communicate with the target host based on the MAC address. In case a match is not found, the target host considers the NA-match message illegitimate, and the message is discarded. Using the AR-match technique secures a target IP address during the entire AR process and thus prevents malicious nodes from disrupting the process. Therefore, DoS on AR process can be fully prevented.

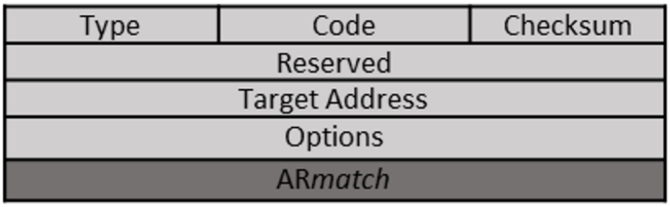


Fig. 5. Message Format of AR-match Technique.

5.2 Workflow of AR-match Technique

The AR-match technique introduces a new secure option called AR*match*, which is appended to each NS and NA message for them to become NS- and NA-match message, respectively. Each message without this option should be silently discarded. The validation is performed on both sides (receiver and sender) to provide additional security by distinguishing between legitimate and illegitimate messages. The workflow of the proposed technique is summarised in the following steps:

- When node A on IPv6 link-local network wants to communicate with the other neighbor node (node B) on the same link, the former should have the latter's MAC address. Firstly, node A must find node B's MAC address in the neighbor cache table. If node B's MAC address is not found, then node A should perform AR-match technique by sending NS-match message asking node B for its MAC address through NA-match message.
- To send the NS-match message, the hash function should be applied on the first 40 and 64 bits of the node B portion of IP address, and the output is called hash value. The hash value of the 40 bits is inserted into AR*match* option, whereas that of the 64 bits is saved into the registry cache for NA message verification later.
- The AR*match* option is attached to the NS message for it to become an NS-match message which is then sent to the SNMA address based on the last 24 bits of the node B IP address (FF02::1:FF00:0/104).
- All existing hosts on the same IPv6 link that have the same SNMA address receive the NS-match message.
- The existing hosts perform a message validation check to ensure the NS-match message comes from a legitimate host based on the existing AR*match* options. IF

the AR*match* option is not found, the receiver (existing hosts) should discard the message.

- Node B will receive the match because its 40-bit hash value will match the 40-bit hash value in AR*match* option. Upon successful matching of hash values, node A asks for node B's MAC address. Node B should hash the 64 bits of node B's IP address, insert the hash value into AR*match*, and append the option into NA message for it to become NA-match message. The NA-match message should be sent to node A as a reply to NS-match message holding node B's MAC address.
- Node A receives the NA-match message and should perform verification by matching the 64-bit hash value. If a match is found, then node A can update its neighbor cache table and start to communicate with the node B based on the MAC address. However, if a match is not found, node A should discard the message and consider the NA-match message comes from illegitimate node.

In this method, a successful AR process can be accomplished in IPv6 link-local network because all nodes can verify the NS and NA messages sent between them during the AR process. Thus, the node can update its neighbor cache table with correct MAC addresses. Fig. 6 shows the workflow when node A (sender) asks for node B's (receiver) MAC address to perform the AR-match technique process.

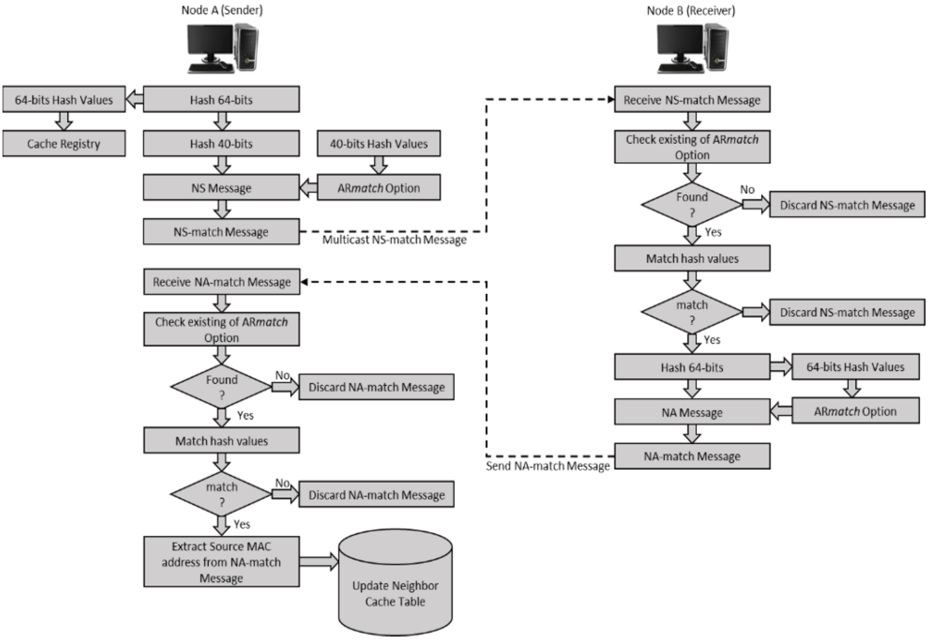


Fig. 6. Workflow of AR-match Technique.

6 Expected Result and Future work

AR-match security technique is proposed to prevent DoS attacks during AR process in IPv6 link-local network. This technique aims to overcome the limitations of the current mechanisms and improve the prevention of DoS-on-AR attacks in terms of processing time and complexity. The AR-match technique based on hash function aims to prevent any attacker from disrupting the process. Any node can verify the NS and NA messages based on the matching of hash values, thereby preventing malicious hosts from disturbing the verification process. As a result, nodes can keep updating their neighbor cache table with a valid MAC address with a less complexity and calculation. The next step is to find the strongest hash function algorithm for high-security purpose, implement the proposed technique and evaluate the results compared with existing related works.

7 Conclusion

DoS attack on AR process is a genuine danger in IPv6 link-local network because all local nodes must know each other's MAC address for a local communication. In the standard AR process, NA and NA messages are unsecure by design, and the target address is revealed, permitting all nodes on the same link to know the target address which is needed to determine its MAC address. Therefore, malicious nodes can forge replies to launch DoS attacks. The proposed AR-match technique aims to hide the target address during AR process through using hash function and thus provide sufficient verification and prevent malicious nodes from sending fake reply and securing IPv6 link-local network from DoS attacks.

References

1. A. Al-Ani, M. Anbar, S. Manickam, … A. A.-A.-P. of the 2017, and U. 2017, "Proposed DAD-match Security Technique based on Hash Function to Secure Duplicate Address Detection in IPv6 Link-local Network," *dl.acm.org*, 2017.
2. T. Narten, W. A. Simpson, E. Nordmark, and H. Soliman, "Neighbor discovery for IP version 6 (IPv6)," 2007.
3. G. Song and Z. Ji, "Anonymous-address-resolution model," *Front. Inf. Technol. Electron. Eng.*, vol. 17, no. 10, pp. 1044–1055, 2016.
4. S. Guangjia, J. Z.-I. J. of S. and Its, and undefined 2015, "Review of Security Research on Address Resolution Protocols," *earticle.net*.
5. T. Narten, S. Thomson, and T. Jinmei, "IPv6 stateless address autoconfiguration," 2007.
6. A. K. Al-Ani, M. Anbar, S. Manickam, A. Al-Ani, and Y.-B. Leau, "Proposed DAD-match Mechanism for Securing Duplicate Address Detection Process in IPv6 Link-Local Network Based on Symmetric-Key Algorithm," 2018, pp. 108–118.
7. W. Simpson, "Neighbor Discovery for IP version 6 (IPv6)," pp. 1–97, 2007.
8. D. Plummer, "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48. bit Ethernet address for transmission on Ethernet hardware," 1982.

9. G. Song, Z. J.-F. of I. T. & Electronic, and undefined 2016, “Anonymous-address-resolution model,” *Springer*.
10. B. Alzahrani, M. Reed, V. V.-P. of the Eleventh, and undefined 2015, “Resistance against brute-force attacks on stateless forwarding in information centric networking,” *dl.acm.org*.
11. O. E. Elejla, M. Anbar, and B. Belaton, “ICMPv6-Based DoS and DDoS Attacks and Defense Mechanisms: Review,” *IETE Tech. Rev.*, vol. 4602, no. August, pp. 1–18, 2016.
12. S. Guangjia, J. Zhenzhou, W. H.-I. J. of Future, and undefined 2015, “A Reverse Address Resolution Process with Variable Length Prefix,” *earticle.net*.
13. N. Kumar, G. Bansal, S. Biswas, and S. Nandi, “Host based IDS for NDP related attacks: NS and NA Spoofing,” in *India Conference (INDICON), 2013 Annual IEEE*, 2013, pp. 1–6.
14. G. Bansal, N. Kumar, S. Nandi, and S. Biswas, “Detection of NDP based attacks using MLD,” in *Proceedings of the Fifth International Conference on Security of Information and Networks*, 2012, pp. 163–167.
15. F. A. Barbhuiya, G. Bansal, N. Kumar, S. Biswas, and S. Nandi, “Detection of neighbor discovery protocol based attacks in IPv6 network,” *Netw. Sci.*, vol. 2, no. 3–4, pp. 91–113, 2013.
16. F. Najjar, M. M. Kadhum, and H. El-Taj, “Detecting Neighbor Discovery Protocol-Based Flooding Attack Using Machine Learning Techniques,” in *Advances in Machine Learning and Signal Processing*, Springer, 2016, pp. 129–139.
17. B. Stockebrand, “Ip security (ipsec),” *IPv6 Pract. A Unixer’s Guid. to Next Gener. Internet*, pp. 311–317, 2007.
18. K. Seo and S. Kent, “Security architecture for the internet protocol,” 2005.
19. J. Arkko, J. Kempf, B. Zill, and P. Nikander, “Secure neighbor discovery (SEND),” 2005.
20. S. Praptodiyono, I. H. Hasbullah, M. M. Kadhum, C. Y. Wey, R. K. Murugesan, and A. Osman, “Securing Duplicate Address Detection on IPv6 Using Distributed Trust Mechanism.,” *Int. J. Simulation--Systems, Sci. Technol.*, vol. 17, no. 26, 2016.
21. Y. Lu, M. Wang, and P. Huang, “An SDN-based Authentication Mechanism for Securing Neighbor Discovery Protocol in IPv6,” *Hindawi Secur. Commun. Networks*, vol. 2017, p. 9, 2017.
22. A. AlSa’deh, H. Rafiee, and C. Meinel, “IPv6 stateless address autoconfiguration: balancing between security, privacy and usability,” in *International Symposium on Foundations and Practice of Security*, 2012, pp. 149–161.
23. J. Wu, G. Ren, and X. Li, “Source address validation: Architecture and protocol design,” in *Network Protocols, 2007. ICNP 2007. IEEE International Conference on*, 2007, pp. 276–283.
24. J. Wu, J. Bi, M. Bagnulo, F. Baker, and C. Vogt, “Source address validation improvement (SAVI) framework,” 2013.
25. H. Krawczyk, R. Canetti, and M. Bellare, “HMAC: Keyed-hashing for message authentication,” 1997.
26. D. McPherson, J. Halpern, and F. Baker, “Source address validation improvement (SAVI) threat scope,” 2013.
27. G. Yao, J. Bi, and P. Xiao, “Source address validation solution with OpenFlow/NOX architecture,” in *Network Protocols (ICNP), 2011 19th IEEE International Conference on*, 2011, pp. 7–12.

Hybridizing Entropy Based Mechanism with Adaptive Threshold Algorithm to Detect RA Flooding Attack in IPv6 Networks

Syafiq Bin Ibrahim Shah[1], Mohammed Anbar[1], Ayman Al-Ani[1], and Ahmed K. Al-Ani[1]

[1] National advanced IPv6 Centre, Universiti Sains Malaysia, Penang, Malaysia

syafiqibrahimshah@usm.my, anbar@nav6.usm.my,
ayman@nav6.usm.my,and ahmedkhallel91@nav6.usm.my

Abstract. The implementation of the neighbor discovery protocol has introduced new security vulnerabilities to Internet protocol version 6 (IPv6) networks. One of the most common attacks being attributed to the IPv6 network layer is the denial of service (DoS) router advertisement (RA) flooding attack. An attacker can flood massive amounts of RA packets to the IPv6 multicast address which cause the hosts inside the link-local network to run out of central processing unit resources due to packet processing overhead. This research proposes a hybrid approach of entropy-based technique combined with the adaptive threshold algorithm to detect the aforementioned attack. By dynamically adapting the threshold and choosing the right entropy feature, the proposed technique is able to detect various scenarios of DoS RA flooding attack, including evasion techniques used by attackers. The proposed technique yields 98% detection accuracy according to the experiment conducted..

Keywords: IPv6, NDP, DoS, Flooding Attack, Entropy, EWMA, RA, Adaptive Threshold Algorithm, Network Security

1 Introduction

Internet service providers across the world are currently and severely curbed by Internet protocol version 4 (IPv4) addresses. The exhaustion of Internet protocol (IP) addresses has served as the main motivation of Internet protocol version 6 (IPv6) development and deployment [1]. IPv6 provides new features, such as stateless address auto configuration (SLAAC), neighbor unreachability detection and duplicate address detection, and the capability to realise end-to-end connectivity without residing on a network address translation (NAT) infrastructure [2]. These features rely heavily on the newly introduced neighbor discovery protocol (NDP).

When a denial of service (DoS) is launched, an attacker typically attempts to consume the network bandwidth and central processing unit (CPU) resources of the target victim. This type of attack can be generated in an IPv6 network by simply flooding

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_31

the link-local network with massive amounts of neighbor discovery (ND) packets. Furthermore, network congestion is bound to happen due to the massive amount of bandwidth consumption in the network. IPv6 inherits the same foundation as that of IPv4, which means that the former has essentially adopted the latter's attributes and qualities, and even shortcomings.

1.1 Neighbor Discovery Protocol (NDP)

NDP operates on layer 2 of the IP stack, and it runs on top of Internet control message protocol version 6 (ICMPv6) [3].

Table 1. Types of NDP messages

Type	Function
Router Solicitation (RS)	Used by hosts for router discovery.
Router Advertisement (RA)	Used by routers to advertise addressing information.
Neighbor Solicitation (NS)	Used for DAD and NUD.
Neighbor Advertisement (NS)	Used by hosts for address resolution.
Redirect	Used to redirect packets from one router to another.

NDP consists of five message types: router solicitation, router advertisement (RA), neighbor solicitation, neighbor advertisement and redirect message. The hosts in an IPv6 network communicate by exchanging these messages. Table 1 lists the types and function of each message [4, 5].

1.2 Denial of Service Router Advertisement Flooding Attack

Routers share information with hosts residing in the same network segment by means of RA message exchange. An RA message contains information, such as network prefix and routing information that can be used by hosts. By default, the hosts in an IPv6 link-local network does not authenticate ingress or egress of IPv6 RA messages. On this basis, a malicious router can spoof and imitate a link-local default gateway and then send crafted messages or flood packets to cause congestion. A hacker can flood the link-local network with crafted address prefix information and default route information. The SLAAC is enabled by default, and thus, the flooded bogus packets will force victims to continuously update their network information. This scenario will lead the victim to utilise its CPU resources until the system ultimately becomes unresponsive [6].

1.3 Types of DoS RA Flooding Attack

A DoS RA flooding attack can be categorised into three types: a default attack, an attack that utilises IPv6 extension headers and an attack involving the fragmentation of packets into smaller pieces or fragments.

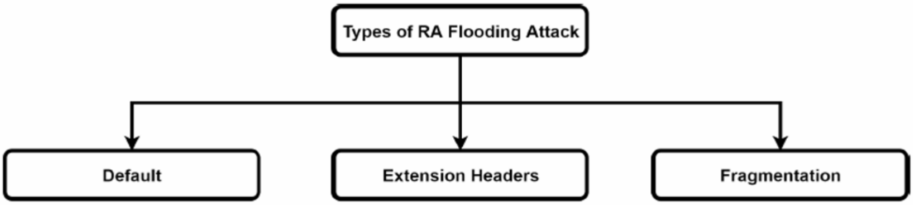


Fig. 1. Types of DoS RA Flooding Attack

The most common type of attack is the default attack in which an attacker floods the network with bogus address prefixes and fake default route information. This type of attack can be detected by the network intrusion detection system, after which an RA guard that can drop the malicious packets is implemented [8], [9].

However, attackers have developed evasion techniques to bypass security validation by forging fake IP extension headers. A security mechanism that only dissects and verifies the initial upper layer of the packet instead of analysing the entire header chain will fail to detect the ulterior malicious content [7].

If a packet is too large to be sent through the network, then the source host will divide the packets into fragments by using fragment headers. In IPv6, the reassembly of these fragmented packets can only be rendered at the destination host. Attackers have exploited these features by fragmenting malicious packets into smaller chunks. However, because the malicious packets cannot be reassembled into their full forms, the security mechanism will defer from performing an analysis, thereby allowing the malicious packets to penetrate the system [8].

2 Related Work

NDP messages can be protected with an Internet protocol security (IPsec) authentication header (AH) to ensure congeniality and integrity of information. The host can verify if the received ND messages are valid and authorised depending on the set of AH security associations (SA), a feature that is helpful in securing the stateless nature of NDPs. Incidentally, SAs that are developed with Internet key exchange version 2 require functional IP addresses, which consequently results in bootstrapping problems. Hence, manual configuration must be performed to secure the NDP with IPSec, but this process is tedious and the configuration is not scalable [10]. The experiments conducted by Xinyu, Ting and Yi [11] proved that IPSec seems powerless if an attacker launches a DoS attack using a legitimate non-spoofed IP address.

An RA guard, a mechanism to prevent the exploitation of RA messages, can be installed on managed switches, but it only forwards the packets received from a port that is known to be connected to an authorised router. Nonetheless, malicious hosts can be prevented from launching DoS attacks [12]. However, although RA guards can effectively verify source legitimacy, the attackers can still circumvent and evade this layer of protection either by forging fake extension headers or by forcing packet fragmentation [7]. Moreover, RA guards can only be installed on specific routers or managed switches, which hinders their portability and scalability.

Aleesa, Hassan and Kamal [13] proposed a detection system for DoS RA flooding by using a rule-based technique. A Web application was used as an attack target, and the application relied on a real IPv6 network that consisted of four routers and six machines. Two rules were defined to detect the RA flooding attack. The first rule is the one-way connection density ratio, while the second rule is the incoming and outgoing ratio of the RA messages. The drawback of the method was the use of a static predefined threshold. In addition, the detection system failed to validate its effectiveness against the improvised evasion techniques attack.

Oshima, Nakashima and Sueyoshi [14] proposed and evaluated an early packet flooding detection mechanism by using the window sizes of 50 packets for DDoS attack and 500 packets for DoS attacks. Source IP and destination port number were used to calculate the entropy value, and the results were used determine whether intrusion occurred or not. An accurate detection result was obtained with a false-negative rate of 5% only during the evaluation phase.

Mousavi and St. Hilaire [15] evaluated the effectiveness of the entropy-based approach to detect DDoS in a software-defined networking environment. The entropy-based approach was selected due to its minimal resource usage and effectiveness against measure randomness. A window size of 50 packets was used for fast detection. The entropy of each window was calculated and compared with an experimental threshold. When the entropy was lower than the threshold, an attack was detected. The detection mechanism was able to detect 96% of the attacks.

3 Detection Technique

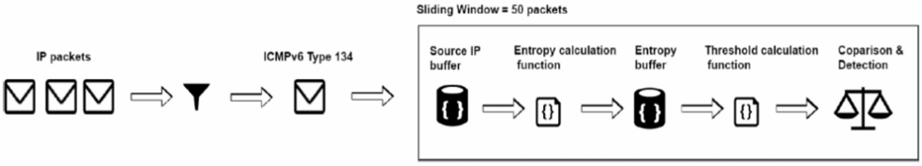

Sliding Window = 50 packets
IP packets
ICMPv6 Type 134
Source IP buffer
Entropy calculation function
Entropy buffer
Threshold calculation function
Coparison & Detection

Fig. 2. Detection Technique

Ingress packets from the network are captured by the traffic capture engine, after which the packets are stored as potential datasets. These datasets are then preprocessed and filtered to reduce irrelevant and unwanted data from the sample. The goal is to reduce the number of relevant traffic features without negatively affecting classification and subsequently improve detection efficiency.

The incoming packets in the network are processed as follows. Only IPv6 packets are filtered in and processed, whereas all other IP packets are filtered out and dropped. The filtration process continues to filter only the ICMPv6 packets. Then, the system verifies whether the packet is a type of ICMPv6 RA message. Type 134 messages are saved as datasets for further computation.

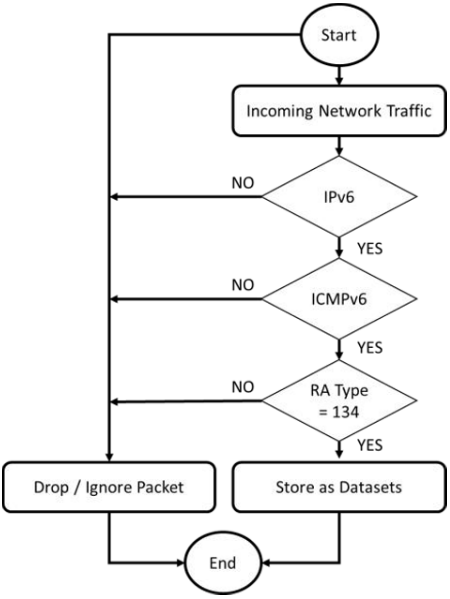


Fig. 3. Packet filtering flowchart

Two essential variables are used to detect network intrusion in the entropy-based approach. The first variable is sliding window size and the second one is threshold value. Sliding window size can be defined based on either a time-period or a number of packet count. The entropy within the sliding window is calculated to measure randomness, an indication of anomalous behaviour. This research proposes a window size of 50 packets to gather statistics. The main reason for choosing 50 packets is due to the small size window. Moreover, a list of 50 values can be calculated much faster than 500 values. Consequently, the intrusion can be detected much earlier because less computation is needed.

Entropy is a common method of network intrusion detection. In this research, entropy is mainly selected because it can measure randomness in the packets that are forwarded into a network. Furthermore, entropy provides a relatively low computation overhead. The higher the randomness of an information given, the higher the entropy and vice versa. When an attack occurs, the entropy value of the source IP address increases because the attacker will always spoof its source IP address on each packet transmission. Hence, a continuously changing source IP is a strong indicator of a DoS RA flooding attack. The Mathematical Theory of Communication, a paper published by C. E. Shannon, defines entropy as:

$$H(x) = -\sum_{i=1}^{n} pi \log pi \qquad (1)$$

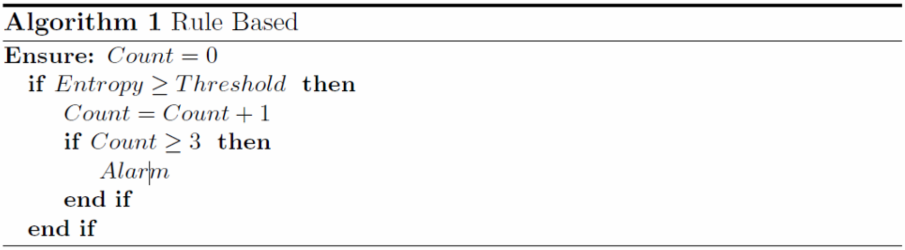
The calculation function computes the entropy of the source IP address of the RA message every 50 packets. A specific threshold value is therefore important in accurately detecting a flooding attack. Ideally, the threshold value should be updated adaptively according to the network traffic condition to perform a much more accurate detection. The adaptive threshold algorithm, a simple algorithm to detect anomalies based on threshold violations over a given interval, can be calculated as:

$$Xn = (\alpha + 1) \ \cdot \mu_{n-1} \qquad (2)$$

In accordance with network traffic changes, the threshold is set adaptively based on the mean value of recent traffic measurement [16]. This research applies the exponential weighted moving average (EWMA) to calculate the mean entropy value in Equation (3).

$$\mu_{n} = \lambda \cdot \mu_{n} + (1 - \lambda) \cdot Xn \qquad (3)$$

Similar to the entropy computation, the mean value of entropy is calculated on each sliding window by using EWMA. The threshold calculation function computes and sets the adaptive threshold based on the most recent mean value calculated. Then, the output of the entropy calculation function is compared with the adaptive threshold values.



Algorithm 1 Rule Based

```
Ensure: $Count = 0$
  if $Entropy \geq Threshold$ then
      $Count = Count + 1$
      if $Count \geq 3$ then
          $Alarm$
      end if
  end if
```

If the entropy value of the source IP address exceeds the threshold value consecutively three times, then a DoS RA flooding attack is likely happening and this will trigger an alarm. In the event that an alarm is triggered, the threshold value is levelled off until the entropy value falls below the threshold value. The reason for this behaviour is to reduce the number of false negative detection.

4 Discussion

In this research, detection accuracy is the metric used to evaluate the effectiveness of the technique. Detection accuracy reflects the capability of an IDS detection engine to accurately determine alerts and generate corresponding alarms in the event of network intrusions. The number of false alarms produced by the system and the percentage of detection are the main indicators of IDS performance

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \qquad (4)$$

Table 2. Definition of abbreviation for accuracy calculation

Abbreviation	Description
TP	True Positive; situation where anomalous events are truly detected as anomaly.
TN	True Negative; situation where normal events are truly recorded as normal.
FP	False Positive; situation where normal events are detected as anomaly.
FN	False Negative; situation where anomalous events are recorded as normal.

Four scenarios are considered and taken as baselines to evaluate the accuracy of the proposed hybrid detection technique. The first scenario involves an attack in the default mode. The second scenario involves an attacker using the extension header to evade the RA guard. The third scenario involves an attack that fragments its packets to avoid detection. The last scenario involves normal IPv6 traffic.

The detection technique is evaluated for each type of dataset (low-, medium- and high-generated packet traffic datasets) and repeated three times to ensure maximal robustness. The results are averaged to obtain the precise accuracy and ensure consistency.

Table 3. Results using scenario 1 datasets

Run	Number of RA Packets	*FP*	*FN*	*TP*	*TN*	Accuracy
1	1451	0	0	30	10	100%
2	3572	0	1	68	16	98.8%
3	7981	0	1	102	2	99%

$$\frac{100+98.8+99}{3} * 100 = 99.26\% \qquad (5)$$

The three runs have yielded an average detection accuracy of 99.26%. The results prove that the detection technique can accurately detect DoS RA flooding attacks, i.e. the attacker floods the link-local network with bogus prefix address and fake router information.

Table 4. Results using scenario 2 datasets

Run	Number of RA Packets	*FP*	*FN*	*TP*	*TN*	Accuracy
1	2186	0	2	48	37	97.7%
2	3322	0	1	64	11	98.6%
3	7343	0	2	141	11	98.7%

$$\frac{97.7+98.6+98.7}{3} * 100 = 98.3\% \qquad (6)$$

The result of 98.3% proves that the detection technique can accurately detect DoS RA flooding attacks, i.e. the attacker floods the link-local network with bogus address prefixes and fake router information that sits below an additional layer of extension headers (Hop_by_Hop) to evade security devices, such as RA guards.

Table 5. Results using Scenario 3 datasets

Run	Number of RA Packets	*FP*	*FN*	*TP*	*TN*	Accuracy
1	1591	0	0	140	17	100%
2	2939	0	0	326	1	100%
3	4720	0	0	481	12	100%

$$\frac{100+100+100}{3} * 100 = 100\% \qquad (7)$$

The result is 100%, which clearly proves that the detection technique can detect DoS RA flooding attacks, i.e. the attacker floods the link-local network with fragmented RA packets to evade security devices, such as RA guards.

Table 6. Results using Scenario 4 datasets

Run	Number of RA Packets	*FP*	*FN*	*TP*	*TN*	Accuracy
1	2856	0	0	0	127	100%
2	2856	0	0	0	127	100%
3	2856	0	0	0	127	100%

$$\frac{100+100+100}{3} * 100 = 100\% \qquad (8)$$

The result is 100%, which clearly proves that the solution does not trigger any false alarm in normal IPv6 traffic conditions.

5 Conclusion and Future Direction

The experimental results prove the capability of the proposed technique to detect RA DoS flooding attacks in IPv6 link-local networks. The effectiveness of the detection technique is evaluated on datasets that have been generated in real time in controlled environments. The results further show that the detection accuracy is high at 98%.

The proposed detection technique can be extended to several promising avenues and directions, such as when building frameworks, to detect NDP flooding attacks in IPv6 link-local networks. Apart from RA DoS flooding, the other types of NDP flooding attacks that may benefit from this research are RS, NS, NA, redirect and MLD flooding attacks. Such directions may be helpful in overcoming the limited attacks that can be detected by the proposed technique in this research. The proposed hybrid detection technique can only detect RA DoS flooding attack in an IPv6 link local network. In future work, the proposed detection technique may be extended for other types of NDP flooding attack detection. The datasets may also be expanded by including more attack scenarios to cover all NDP flooding types.

References

1. Al-Ani, A.K., Anbar, M., Manickam, S., Al-Ani, A., Leau, Y.-B.: Proposed DAD-match Mechanism for Securing Duplicate Address Detection Process in IPv6 Link-Local Network Based on Symmetric-Key Algorithm. In: International Conference on Computational Science and Technology. pp. 108–118 (2017)
2. Graziani, R.: IPv6 fundamentals: a straightforward approach to understanding IPv6. Pearson Education (2012)
3. Al-Ani, A.K., Anbar, M., Manickam, S., Al-Ani, A., Leau, Y.-B.: Proposed DAD-match Security Technique based on Hash Function to Secure Duplicate Address Detection in IPv6 Link-local Network. In: Proceedings of the 2017 International Conference on Information Technology. pp. 175–179 (2017)
4. Anbar, M., Abdullah, R., Saad, R., Hasbullah, I.H.: Review of Preventive Security Mechanisms for Neighbour Discovery Protocol. Adv. Sci. Lett. 23, 1130611310 (2017)
5. Anbar, M., Abdullah, R., Al-Tamimi, B.N., Hussain, A.: A Machine Learning Approach to Detect Router Advertisement Flooding Attacks in Next-Generation IPv6 Networks. Cognit. Comput. 114 (2017)
6. Elejla, O.E., Belaton, B., Anbar, M., Alnajjar, A.: Intrusion detection systems of ICMPv6-based DDoS attacks. Neural Comput. Appl. 112 (2016)
7. Gont, F.: Implementation Advice for IPv6 Router Advertisement Guard (RA-Guard). (2014)
8. Levy-Abegnoli, E., de Velde, G., Popoviciu, C., Mohacsi, J.: IPv6 router advertisement guard. (2011)
9. M. Anbar, R. Abdullah, B. Al-Tamimi, A. H.-C. Computation, and undefined 2017, “A Machine Learning Approach to Detect Router Advertisement Flooding Attacks in Next-Generation IPv6 Networks,” Springer.
10. Arkko, J., Aura, T., Kempf, J., Mntyl, V.-M., Nikander, P., Roe, M.: Securing IPv6 neighbor and router discovery. In: Proceedings of the 1st ACM workshop on Wireless security. pp. 7786 (2002)
11. Yang, X., Ma, T., Shi, Y.: Typical dos/ddos threats under ipv6. In: Computing in the Global Information Technology, 2007. ICCGI 2007. International Multi-Conference on. p. 55 (2007)
12. Chown, T., Venaas, S.: Rogue IPv6 Router Advertisement Problem Statement. (2011)
13. Aleesa, A.M., Hassan, R., Kamal, S.U.M.: A rule-based technique to detect router advertisement flooding attack against biobizz web application. Adv. Sci. Lett. 22, 18871891 (2016)
14. Oshima, S., Hirakawa, A., Nakashima, T., Sueyoshi, T.: DoS/DDoS detection scheme using statistical method based on the destination port number. In: Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP09. Fifth International Conference on. pp. 206209 (2009)
15. Mousavi, S.M., St-Hilaire, M.: Early detection of DDoS attacks against SDN controllers. In: Computing, Networking and Communications (ICNC), 2015 International Conference on. pp. 7781 (2015)
16. Cisar, P., Cisar, S.M.: EWMA statistic in adaptive threshold algorithm. In: Intelligent Engineering Systems, 2007. INES 2007. 11th International Conference on. pp.5154 (2007)

Frequent Itemset Mining in High Dimensional Data: A Review

Fatimah Audah Md. Zaki and Nurul Fariza Zulkurnain

Department of Electrical and Computer Engineering,
International Islamic University Malaysia.
fatimah.audah@gmail.com, nurulfariza@iium.edu.my

Abstract. This paper provides a brief overview of the techniques used in frequent itemset mining. It discusses the search strategies used; i.e. depth first vs. breadth-first, and dataset representation; i.e. horizontal vs. vertical representation. In addition, it reviews many techniques used in several algorithms that make frequent itemset mining more efficient. These algorithms are discussed based on the proposed search strategies which include row-enumeration vs. column-enumeration, bottom-up vs. top-down traversal, and a number of new data structures. Finally, the paper reviews on the latest algorithms of colossal frequent itemset/pattern which currently is the most relevant to mining high-dimensional dataset.

Keywords: Data mining, High-dimensional data.

1 Introduction

Data mining is the process to uncover hidden knowledge, identify trends and patterns, and discovering new rules and relationships from large databases [1]. It is regarded as the most important step in a process called knowledge discovery in databases (KDD), which is closely related to another important process; data warehousing, where operational data from a few databases are first cleaned then integrated into a data warehouse. Frequent pattern mining includes four main types including sequential pattern, frequent itemset, and graph mining. Frequent itemset mining is an important types of data mining that was initially developed for market basket analysis [2], which examines customer behaviour and identifies sets of items that often purchased together. This information may be used to maximize organization profit by efficient products arrangement on shelves, deciding related products to be on discounted price, etc. which will encourage customers to spend more. The algorithm will find similar sets of items exist at least a minimum amount of times, and will be represented in the form of association rules. Association rules that satisfy the minimum support and minimum confidence thresholds are considered interesting.

However, mining frequent itemsets from a large data set has a major drawback. The huge number of generated itemsets that satisfies the minimum support will be too much for computer memory to store and make further processing. Therefore, the concepts of closed frequent itemset (CFI) and maximal frequent itemset (MFI) were introduced. CFI is an itemset with support that satisfies the minimum support and does

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_32

not have proper supersets with the same support. On the other hand, MFI is an itemset that satisfies the minimum support and none of its superset satisfies the minimum support. Though MFI is more compact than CFI, it resulted in loss of information. Instead, the whole set of frequent itemset can be derived based on information of the CFI, i.e.; the itemset and its support [3]. Therefore, it can be concluded that CFI contains complete information of frequent itemsets. Therefore, the output of mining algorithm can be reduced by confining it to closed frequent itemset.

2 High-Dimensional Data

Currently, we are in the age of massive automatic data collection and the existing trends are likely to accelerate in the near future. This involves another kind of dataset, known as high-dimensional data, which is categorized by a relatively less number of rows compared to columns (or dimensions) [5]. For instance, many systems-biology technologies in cancer research are used to build better predictive models of diagnosis, therapies, and drugs. These technologies require mining meaningful statistical and biological information from high-dimensional data, wherein each sample is defined by hundreds or thousands of measurements, which are obtained simultaneously [4]. The data are high-dimensional due to the huge number of individual measurements are obtained on each specimen. Other applications that deal with high-dimensional data include ecological data where 200 plots and 150 species totally for sociology analysis, financial data, e.g. stocks exchange prices based on daily, hourly, and minutely, many applications in computer vision e.g., basic facial recognition algorithm, where n images, each with a resolution of m pixels by k pixels and Earth Observation Data/Geographical data may have hundreds of dimensions e.g., longitude, latitude, temperature, pressure, rainfall, humidity, altitude, time, season, soil type, and many more.

3 Frequent Itemset Mining Technologies

The basic algorithms for frequent itemset mining are Apriori, Eclat , and FP-growth. These are characterized based on the ways they traverse the search space, i.e.; breadth-first or depth-first. They also differ in the way the conditional transaction databases are represented, i.e.; horizontal representation or vertical representation.

3.1 Search Strategies

Breadth-First Search

Breadth-first search is used in the Apriori algorithm, implemented with a prefix tree that is built level by level. Employing breadth-first search means that, from the first level, each node will be created and scanned; starting from the leftmost node towards the right. Each level of the tree must be fully explored to discover frequent itemsets before moving onto the next level. For every frequent itemset discovered, a child node will be created, hence a new level is added. Frequent itemsets are discovered by candidate generation with two pruning steps namely a priori pruning and a posteriori pruning. In a priori pruning, all itemsets that possess an infrequent subset are deleted. While in a posteriori pruning, all infrequent item sets are eliminated.

The drawbacks of Apriori algorithm are the big number of candidate itemsets and the need to scan the database repetitively to count the support of the candidates, which reduced its performance, both in speed and memory. Furthermore, in breadth-first search, pointers are stored for every level's child nodes in order to know which nodes to traverse for the next level and this uses a large amount of memory.An algorithm named A-CLOSE [9] was developed based on Apriori framework. It generates frequent itemsets from a database in two phases, i.e. discovers all CFI then generates all subsets and derives their support from the CFI. Discovering CFI has reduced the number of generated candidate itemsets and scanning of database is only needed in the first phase.

Depth-First Search

In FP-growth, the database is scanned to derive a set of frequent items according to their frequency in descending order which are stored in a frequent-pattern tree, or FP-tree. The pattern growth starts from frequent pattern of length-1 pattern, creating the respective conditional pattern based as well as conditional FP-tree.

Mining is performed iteratively with the tree using depth-first search, starting at the root and following one of the branches of the tree up until the node with no children is found, and continue the search at the previous node with unexplored children. FP-growth has change the method into concatenating the suffix of shorter frequent patterns, instead of finding long frequent patterns [6]. Unlike breadth-first search, it is not necessary to store all child pointers at every level, hence requires much lower memory.FP-growth based has been the most approached method by researchers as it compressed large databases into a compact FP-tree which eliminates candidate generation and avoids repeated scanning of the database. Several CFI mining algorithms have been developed as an extensions to the FP-growth such as CLOSET [10], CLOSET+ [11], FP-Close [12], CHARM [13], and AFOPT [14].

3.2 Data Representation

Horizontal Dataset Format

Apriori and FP-growth mine frequent patterns from horizontal dataset, which is the normal way of storing databases. In a horizontal dataset, the transactions are listed as the rows which lists the items as the columns.

Vertical Data Format

A vertical data format is an inverted version of the original dataset and is first proposed in the Eclat algorithm [7]. The database is scanned and for each item, the transactions id that contains the items is listed. The frequent itemset is generated starting from a single item, like Apriori, and employs depth-first search as FP-growth. The advantage of this representation is that each k-itemset has complete information to count the support. Therefore, it is unnecessary to find out the support of the (k+1)-itemset from the dataset.

Obviously, the mining time taken by Eclat depends on the length of the transaction items list, hence taking less time for shorter list [8]. CHARM is an example of algorithm that discovers CFI using vertical data format and shown to have perform better than A-CLOSE and CLOSET [13]. Meanwhile, MAFIA [15] uses vertical bitmaps to store the database, which is ideal to generate an itemset and counting its support. CARPENTER [16] has been proposed to mine CFI in long biological datasets using vertical data format and pattern-growth technique and shown to outperform CHARM and CLOSET.

4 Previous Works on Frequent Itemset Mining

This section discussed the previous works done on FIM from the year 1998 to 2013. The works are categorized based on whether their approach is to mine MFI or CFI. Apart from that, the algorithms can be distinguished into column- and row-enumeration.

Dataset with less number of columns compared to rows can be efficiently enumerated according to its column as the number of column combinations is smaller than row combination. In contrast, row-enumeration is more suitable for datasets with larger number of columns; i.e. high dimensional data. Both column and row-enumeration approach exploit either bottom-up or top-down search strategy. In bottom-up strategy, the dataset is searched starting from the smallest itemset/rowset, while top-down strategy starts from the largest itemset/rowset.

4.1 Mining Closed Frequent Itemset (CFI)

To improve memory consumption, CLOSET+ [32] used a hybrid tree-projection method. A depth-first search is performed on the horizontal data format while the local frequent items is generated by constructing a projected database. The first algorithm to enumerate the rows is CARPENTER [33] which also uses vertical database. In CARPENTER, conditional transposed tables are repeatedly generated, while traversing the row-enumeration tree in a depth-first manner, resulted to an improved mining time. Many of the proposed row-enumeration algorithms had their foundation based on CARPENTER.

FARMER [34] is an algorithm specifically developed for mining interesting rule groups from microarray datasets, i.e. a grid of DNA segment and an example of high dimensional dataset. This algorithm is enhanced with efficient search pruning strategies based on user-specified thresholds (minimum support, minimum confidence and minimum chi-square value). Like FARMER, TOPKRGS [35] row-enumerate the rule groups but differs in differentiating significant rules by using preference selection (top-k). In addition, its efficiency is improved with the use of compact prefix-tree. COBBLER [36] is an algorithm that mined CFI and able to switch between row and column enumeration while mining. The switching is determined from the enumeration cost of the subtrees which helps in dealing with different kind of dataset. An alternative mining task, TFP algorithm [37] was developed for mining CFI without having to specify for a minimum support (min_supp). The idea is to mine a desired number of CFI of length no less than a specified minimal length (min_l) of each closed itemset. Based on specific applications, the value is easily specified compared to min_supp which requires detailed information on mining query and the types of data. Performance studies show that even with the best adjusted min_supp, this algorithm outperforms CHARM and CLOSET+.

The previous algorithms discussed; CARPENTER, COBBLER, FARMER, and TOPKRGS, used bottom-up search strategy to search through the row-enumeration tree. However, the minimum support threshold cannot be fully used to prune search space, thus the time and memory are not efficient. Therefore, the row-enumeration tree searched with top-down strategy is proposed in TD-Close [38]. This make pruning using minimum support threshold fully utilized, hence smaller search space. An effective closure-checking strategy that avoids multiple scanning of the database is also proposed. An improved version of TD-Close, i.e. TTD-Close [39] use new pruning strategies and new data structure resulted in faster mining process.

A mining method to describe interesting gene relationships from microarray datasets is presented in MAXCONF [40], which addresses the issue of pattern explosion due to support-based, row-enumeration method. Instead of using min_supp, confidence measure is used to prune the search space. A comparative analysis using several microarray datasets revealed that MAXCONF beats support-based rule mining in terms of scalability and rule extraction.

5 Mining Colossal Itemset

The concepts of CFI and MFI are introduced to limit the number of generated itemsets from mining frequent itemsets from large databases. However, some mining task such as analysis of microarray data in bioinformatics still produced a huge number of small frequent patterns. Unfortunately, in practice, patterns that are larger in pattern size give more important meaning than shorter ones. These large patterns are called colossal patterns [41]. Hence, it is ineffective to execute a mining algorithm that will find all sizes of frequent patterns when the ones needed are only the colossal ones.

The concept of colossal pattern was first introduced in an algorithm named Pattern-Fusion [41] that finds an estimation of colossal patterns. Comparing this algorithm with Apriori and FP-growth which adopt pattern-growth mining strategies, Pattern-Fusion merged the core-patterns to discover the colossal pattern, hence eliminates the need to scan a big number of small to medium sized patterns and avoids exhaustive level-wise pattern tree traversal. In Pattern-Fusion, the colossal pattern is discovered by combining randomly selected small patterns and the support is counted by individual database scan. On the other hand, Colossal Pattern Miner (CPM) [42] suggested another way of selecting the small patterns which will be merged. The method includes segregating the small patterns of redundant colossal patterns based on the frequencies, which eliminates non-colossal patterns. Furthermore, vertical data format is used so as repeated scanning of the database can be avoided. Unfortunately, there is no study conducted to compare the performance of this algorithm with Pattern-Fusion.

BVBUC algorithm [43] represents the dataset with a bit matrix (which compresses the dataset), such that when a row contain a specific item, then the bit of the row bit string will be set to 1 while others are 0. A bottom-up row-enumeration tree is built until min_supp as the largest pattern of each tree branches is generated minsup level. Most pattern mining algorithms rely on the min_supp as the search constraint, and lower min_supp leads to the discovery of many frequent patterns, resulted in increased run-time and insufficient memory due to the increase size of the search tree. However, in BVBUC, lower min_supp leads to lower number of level to be created in the tree. This is achieved by pruning branches that do not satisfy min_supp. Another pruning method is the minimum acceptable number of items in a colossal pattern, where the process of growing a branch is stopped once the the number of itemsets is lower than the minimum threshold. Since the algorithm applies bottom-up row-enumeration search, the first mined itemset is the biggest one, thus increasing the mining speed. These two pruning methods lead to much efficiency improvement in the mining task. In addition, closedness-checking strategy is done by performing AND operation of an itemset bit matrix to each items column bit matrix and determine whether they satisfies the minimum support threshold. Experiments have been conducted using real standard datasets and microarray datasets and results show BVBUC outperforms both CPM and Pattern-Fusion.

An algorithm termed DisClose [5] extracts colossal closed itemsets by first enumerating large cardinality itemsets, and then builds smaller itemsets. As BVBUC, this is done by traversing the row-enumeration tree in bottom-up way. The algorithm begins by transposing the dataset and stores the itemsets in Compact Row-Tree (CR-Tree) which is used during the search process. Several rowset values which represent an itemset that satisfies the minimum number of items, mincard, will share the same node, making the proposed tree compact. Itemsets that do not meet minimum mincard are not added as they will not contribute to finding colossal (large) itemsets. A closedness-checking method based on a unique generator is proposed, where for a closed rowset β = {t1, t2….tk}, all its subsets that do not have the prefix as β can be pruned. This algorithm's performance is evaluated by comparing it with FP-Close, D-Miner, CARPENTER and TTD-Close, and results show that it is scalable and can efficiently extract colossal closed itemsets even for low support threshold.

A recently proposed algorithm which applies for biological datasets, Doubleton Pattern Mining (DPM) [44] represented the dataset vertically and used vector intersection operator to find colossal patterns. The algorithm proposed a data structure named D-struct which stores doubleton data matrix in an array pair. The algorithm consists of two stages; finding doubleton pattern that is stored in D-struct, then enumerate the column by vector intersection to discover colossal patterns. This has resulted to a reduced execution time as repeated database scan in unnecessary. Analysis shows its efficiency in mining several biological data and is better than Colossal Pattern Miner (CPM). An improved algorithm of DPM called DPMine [45] discovers Doubleton Patterns which are then stored in DPT+ tree. The tree is built as top down column enumeration using bitmaps to generate colossal pattern sequence with vector intersection operator. Experimental results on real Lung Cancer and Prostate Cancer dataset show that this approach outperforms CPM and BVBUC.

Recently proposed algorithms named CP-Miner and its improved version; i.e. PCP-Miner [46] has been proposed to mine colossal patterns according to the τ-core ratio. A pattern is considered τ-core if the ratio of its superset's support to its own support is more than the predetermined τ-core ratio. A τ-core pattern is deemed colossal if there is no supersets exist in the database. The difference between these algorithms with BVBUC is that BVBUC does not require the use of τ-core ratio, but a minimum support threshold and minimum acceptable number of items in an itemset, as proposed by DisClose. However, they have been proven to perform better than BVBUC in terms of runtime.

6 Conclusion

Frequent itemset mining has been an important task for discovering useful knowledge for various applications. Throughout the years, FIM algorithms have evolved to become more efficient in terms of run-time, memory consumption, and even usefulness of the extracted itemset. From this review, it can be observed that the trend of research in FIM in high-dimensional data is moving towards mining colossal itemsets/patterns which are critical in many trending applications.

ACKNOWLEDGMENT

The authors would like to thank the Malaysian Government and International Islamic University Malaysia (IIUM) for the research grant under Fundamental Research Grant Scheme (FRGS) Research Project FRGS14-139-0380.

References

1. Adriaans, P. & Zantinge, D. (1996). Data Mining. Harlow, England: Addison Wesley Longman.
2. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in largedatabases. In: Proceedings of the 1993ACM-SIGMOD international conference on managementDFGHJL'of data (SIGMOD'93), Washington, DC, pp 207–216.
3. Han, J., Kamber, M., and Pei, J. (2012). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.
4. Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews. Cancer, 8(1), 37–49. http://doi.org/10.1038/nrc2294
5. Zulkurnain, Nurul Fariza (2012). DisClose: Discovering Colossal Closed Itemsets from High Dimensional Datasets via a Compact Row-Tree.

6. Han, J., Pei, J., and Yin, Y., (2000). Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 1–12.
7. Zaki, M. J., (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3), pp. 372-390.
8. Borgetl, C. (2012), WIREs Data Mining Knowl Discov 2012, 2: 437–456.
9. Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) Discovering frequent closed itemsets for association rules. In: Proceeding of the 7th international conference on database theory (ICDT'99), Jerusalem, Israel, pp 398–416.
10. Pei J, Han J, Mao R (2000) CLOSET: an efficient algorithm for mining frequent closed itemsets.In: Proceeding of the 2000 ACM-SIGMOD international workshop data mining and knowledge discovery (DMKD'00), Dallas, TX, pp 11–20.
11. Wang J, Han J, Pei J (2003) CLOSET+: searching for the best strategies for mining frequent closed itemsets. In: Proceeding of the 2003 ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03), Washington, DC, pp 236–245.
12. Grahne G, Zhu J (2003)Efficiently using prefix-trees in mining frequent itemsets. In: Proceeding of the ICDM'03 international workshop on frequent itemset mining implementations (FIMI'03), Melbourne, FL, pp 123–132.
13. Zaki MJ, Hsiao CJ (2002) CHARM: an efficient algorithm for closed itemset mining. In: Proceeding of the 2002SIAMinternational conference on data mining (SDM'02),Arlington,VA, pp 457–473.
14. Liu G, Lu H, Lou W, Yu JX (2003) On computing, storing and querying frequent patterns. In: Proceeding of the 2003 ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03), Washington, DC, pp 607–612.
15. Burdick D, Calimlim M, Gehrke J (2001) MAFIA: a maximal frequent itemset algorithm for transactional databases. In: Proceeding of the 2001 international conference on data engineering (ICDE'01), Heidelberg, Germany, pp 443–452.
16. Pan F, Cong G, Tung AKH, Yang J, Zaki M (2003) CARPENTER: finding closed patterns in long biological datasets. In: Proceeding of the 2003 ACMSIGKDD international conference on knowledge discovery and data mining (KDD'03),Washington, DC, pp 637–642.
17. R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.
18. R. Agrawal, C. Aggarwal, and V. Prasad, "Depth first generation of long patterns, "In SIGKDD, 2000.
19. D.-I. Lin and Z. M. Kedem, "Pincer-search: an efficient algorithm for discovering the maximum frequent set," Knowledge and Data Engineering, IEEE Transactions on, vol. 14, no. 3, pp. 553–566, 2002.
20. Q. Zou, W. W. Chu, and B. Lu, "SmartMiner: A depth first algorithm guided by tail information for mining maximal frequent itemsets," in Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, 2002, pp. 570–577.
21. Hua-Fu Li, Suh -Yin Lee and Man-Kwan Shan, "Online mining (recently) maximal frequent itemsets over data streams," Research Issues in Data Engineering: Stream Data Mining and Applications, 2005. RIDE-SDMA 2005. 15th International Workshop on , pp: 11-18, 3-4 April 2005.
22. K. Gouda and M. J. Zaki, "Genmax: An efficient algorithm formining maximal frequent itemsets," Data Mining and Knowledge Discovery, vol. 11, no. 3, pp. 223–242, 2005.
23. T. Hu, S. Y. Sung, H. Xiong, and Q. Fu, "Discovery of maximum length frequent itemsets," Information Sciences, vol. 178, no. 1, pp. 69–87, 2008.

24. Tran Anh Tai; Ngo Tuan Phong; Nguyen Kim Anh, "An Efficient Algorithm for Discovering Maximum Length Frequent Itemsets," in Knowledge and Systems Engineering (KSE), 2011 Third International Conference on , vol., no., pp.62-69, 14-17 Oct. 2011.
25. M.Rajalakshmi,Dr.T.Purusothaman, Dr.R.Nedunchezhian, "Maximal Frequent Itemset Generation Using Segmentation Approach", International Journal of Database Management Systems (IJDMS),Vol. 3, No.3, Aug 2011.
26. NVB Gangadhara Rao, Sirisha Aguru, "A Hash based Mining Algorithm for Maximal Frequent Item Sets using Double Hashing," Journal of Advances in Computational Research: An International Journal Vol. 1 No. 1-2 (Jan-Dec, 2012).
27. P.C.S.Nagendra setty, D. Haritha, and Vedula Venkateswara Rao, "Improved Maximal Length Frequent Item Set Mining," International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 2, August 2012.
28. G. Vijay Kumar, Dr. V. Valli Kumari, "MaRFI: Maximal Regular Frequent Itemset Mining using a pair of Transactionids", International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345, Vol. 4, No. 07, Jul 2013.
29. G. Vijay Kumar, V. Valli Kumari, "IncMaRFI: Mining Maximal Regular Frequent Itemsets In Incremental Databases," International Journal of Engineering Science and Technology (IJEST), ISSN: 0975- 5462 Vol. 5, No.08, Aug. 2013.
30. Jnanamurthy HK, Vishesh HV, Vishruth Jain, Preetham Kumar, Radhika M. Pai, "Discovery of Maximal Frequent Item Sets using Subset Creation," International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 3, No. 1, Jan. 2013.
31. Maha Attia Hana, "MVEMFI: Visualizing and Extracting Maximal Frequent Itemsets," Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, pp.183-189, Sep-Oct 2013.
32. Wang J, Han J, Pei J. "CLOSET+: searching for the best strategies for mining frequent closed itemsets," In: Proceedingsof the The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA: ACM Press, 2003;236–45.
33. Pan F, Cong G, Tung AK, et al. "Carpenter: finding closed patterns in long biological datasets," In: Proceedings of the the Ninth SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA: ACM Press, 2003; 637–42.
34. Cong G, Tung AK, Xu X, et al. "FARMER: finding interesting rule groups in microarray datasets," In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Paris, France: ACM Press, 2004; 143–54.
35. Cong G, Tan K, Tung AK, et al. "Mining top-K covering rule groups for gene expression data," In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, USA: ACM Press, 2005; 670–81.
36. F. Pan, A. Tung, G. Cong, and X. Xu, "Cobbler: Combining column and row enumeration for closed pattern discovery," in Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on, 2004, pp. 21–30.
37. J. Wang, J. Han, Y. Lu, and P. Tzvetkov, "TFP: An efficient algorithm for mining top-k frequent closed itemsets," Knowledge and Data Engineering, IEEE Transactions on, vol. 17, no. 5, pp. 652–663, 2005.
38. H. Liu, J. Han, D. Xin, and Z. Shao, "Mining frequent patterns from very high dimensional data: A top-down row enumeration approach," in Proceeding of the 2006 SIAM international conference on data mining (SDM'06), Bethesda, MD, 2006, pp. 280–291.
39. H. Liu, X. Wang, J. He, J. Han, D. Xin, and Z. Shao, "Top-down mining of frequent closed patterns from very high dimensional data," Information Sciences, vol. 179, no. 7, pp. 899–924, 2009.

40. McIntosh T, Chawla S. High confidence rule mining for microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 2007; 4:611–23.
41. Zhu F, Yan X, Han J, et al. Mining colossal frequent patterns by core pattern fusion. In: Proceedings of the IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey: ACM Press, 2007; 706–15.
42. Madhavi D. and Mogalla S., "An Efficient Approach to Colossal Pattern Mining," IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.1, January 2010.
43. Mohammad Karim Sohrabi and Ahmad Abdollahzadeh Barforoush, "Efficient colossal pattern mining in high dimensional datasets," Knowledge-Based Systems, vol. 33, pp. 41–52, 2012.
44. K.Prasanna and M.Seetha, "A Doubleton Pattern Mining Approach for Discovering Colossal Patterns from Biological Dataset," International Journal of Computer Applications, vol. 119, no.21, June 2015.
45. K.Prasanna and M.Seetha, "Efficient and Accurate Discovery of Colossal Pattern Sequences from Biological Datasets: A Doubleton Pattern Mining Strategy (DPMine)," Eleventh International Multi-Conference on Information Processing (IMCIP-2015), Procedia Computer Science, vol. 54, 2015, pp. 412 – 421.

Thanh-Long Nguyen, Bay Vo, and Vaclav Snasel. 2017. Efficient algorithms for mining colossal patterns in high dimensional databases. Know.-Based Syst. 122, C (April 2017), 75-89.

Validation of Bipartite Network Model of Dengue Hotspot Detection in Sarawak

Woon Chee Kok and Jane Labadin

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Malaysia
wooncheee.kok@gmail.com, ljane@unimas.com

Abstract. This paper presents the verification and validation processes in producing a realistic bipartite network model to detect dengue hotspot in Sarawak. Based on the result of previous published work, ranking of location nodes of possible dengue hotspot at Sarawak are used to illustrate the validation by comparing the Spearman rank correlation coefficients (SRCC) between the network models. UCINET 6 is used to generate a benchmark ranking result for model verification. A centrality measure analysis feature available in UCINET is used to determine the node centrality of a network model. The validation results show strong ranking similarity for all three groups of network models with good Spearman rank correlation coefficients values of 1.000, 0.8000 and 0.8824 ($\rho > 0.80$; $p < 0.001$) respectively. The top-ranked locations are seen as dengue hotspots and this study demonstrate a new approach to model dengue transmission at district-level by locating the hotspots and prioritizing the locations according to vector density.

Keywords: Bipartite Network, Dengue Hotspot, Validation, Verification

1 Introduction

Hotspot detection is also referred to as geospatial patterns modeling, area clustering modeling or spatial structure modeling. Typical studies of hotspots detection aim to map the disease occurrences with demographic and environmental properties to better understand the spread of disease. Hotspot detection studies are commonly carried out on infectious diseases such as tuberculosis, dengue and others [1]. Researchers are interested to identify specific environmental properties that are highly correlated to the distribution of the dengue patients in the area with unusually elevated response and thus can be categorized as hotspot.

In a disease domain, it is important for public health authorities to assess whether there is an unusual spatial aggregation of disease cases for them to make decision in allocating resources for prevention. Environmental properties relate to hotspot detection studies in which the physical characteristics of a site and the distribution of the dengue occurrences are included [2].

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_33

Significant environmental properties are identified and were used to quantify the nodes in the network model from our previous study [3]. Altitude, temperature, precipitation and humidity and temperature are identified as important parameters to quantify the link between the nodes. Dengue vectorial capacity (VC) was defined by [6] shown in equation 1 to describe dengue vector's ability to spread the disease among humans and takes into account host, virus and vector interactions where m represents vector density, b represents transmission probability, c represents infection probability, a represents number of bites, p represents vector survival rate and n represents the extrinsic incubation period (EIP). The parameters involved in the equation are significant predictors to describe the abundance and characteristics of the dengue vector at one hotspot from biological perspectives.

$$VC = \frac{mbca^2p^n}{-\ln p} \tag{1}$$

The EIP is the viral incubation period between the time when a mosquito takes a blood meal and the time when that mosquito become infectious [4]. In other words, little time is required for the virus to replicate after the mosquito took a blood meal from an infected person and spread through the body of the mosquito until the virus can be passed on to another host. Knowledge about the EIP of dengue for *Ae. Aegypti* is scarce. Thus, life cycle duration is introduced to examine the risk of dengue occurring in a given area [2,4-5].

The transmission of dengue is strongly dependent on mosquito biting behavior and this biting behavior is temperature-dependent [6-8]. The biting rate can be defined as the number of blood-feeding of a mosquito taken per unit time [9]. Vector biting which is associated with temperature was applied in this study.

Vector survival has long been recognized as one of the most significant parameters in mosquito-borne disease transmission [2,4-5]. Experiments conducted on female *Aedes* mosquitoes over the temperature range between 15 ℃ to 34 ℃ revealed that the survival rate are ranged from 23.5% to 67% per day and this parameter is used in this study [5].

As a summary, life cycle, survival, biting behavior and frequency of mosquito visits the location are possible important model predictors that contribute to dengue spread.

This paper presents our continuous research in applying bipartite network modeling approach to predict dengue hotspot in Sarawak [3]. Model verification and validation process of the bipartite dengue contact (BDC) network is highlighted in this paper.

2 Materials and methods

Model verification is a procedure of determining if the implementation of a model has been done correctly [10]. It is also reported as the process to answer the question of "have we built the model right?" [11]. The verification is conducted through benchmark verification. Benchmark verification aims to verify research procedures by comparing the result obtained from the BDC network model and the result produced by a benchmark system [12]. This can be achieved by calculating the Root Mean Square Error (RMSE) between the ranking values obtained from this research and the values produced by a benchmark system. RMSE is the most popular measure of goodness-of-fit, where a good RMSE value to verify the model is built right when the values is not greater than 0.05 [13-14]. The benchmark verification involved three steps as shown in Fig. 1 which is adopted from previous studies [12,15].

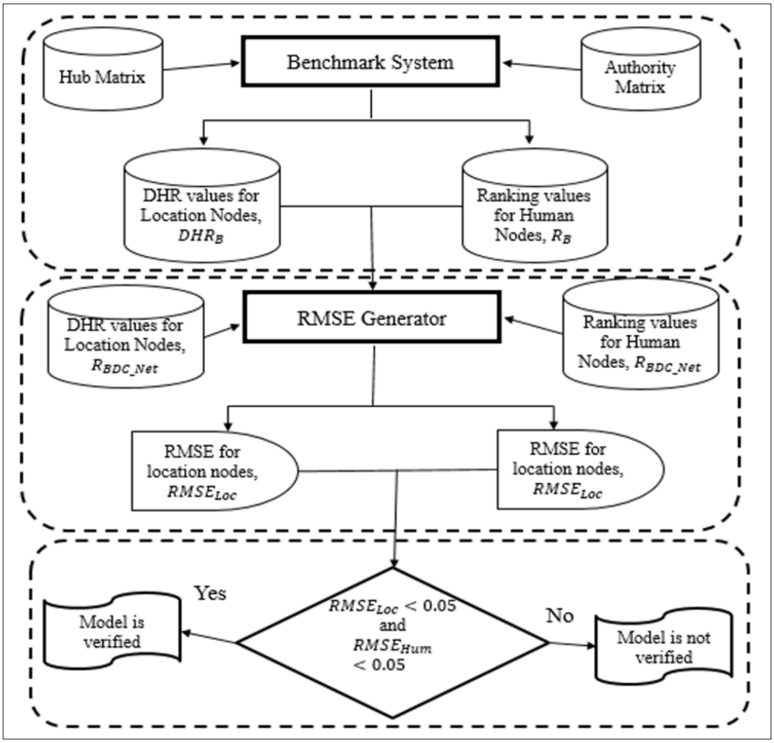

Hub Matrix
Benchmark System
Authority Matrix
DHR values for Location Nodes, DHR_B
Ranking values for Human Nodes, R_B
DHR values for Location Nodes, R_{BDC_Net}
RMSE Generator
Ranking values for Human Nodes, R_{BDC_Net}
RMSE for location nodes, $RMSE_{Loc}$
RMSE for location nodes, $RMSE_{Loc}$
Yes
$RMSE_{Loc} < 0.05$ and $RMSE_{Hum} < 0.05$
No
Model is verified
Model is not verified

Fig. 1. Benchmark Verification Process

UCINET 6 for Windows is a powerful network analysis software and able to handle two million nodes and also two-type-node networks [16]. It supports many network analysis methods, which includes centrality measure, subgroup identification and elementary graph theory. The centrality measure analysis feature available in UCINET can determine the node centrality of a network. The node centrality obtained by measuring the relative importance of a node within the network. Bonacich eigenvector centrality provided in UCINET is one of the centrality measures. This Bonacich eigenvector centrality can identify the node which is more central if they are connected to other nodes that are themselves central. Hence, it relates both the number of connections to other nodes and also measure how well the neighbors are connected.

Firstly, the hub matrix that is generated from the Hypertext Induced Text Search (HITS) algorithm is loaded into UCINET as shown in Fig 1. It is then transformed into UCINET system files, header file (in ##H file) and data file (in ##D file). Next, the header file is used for centrality analysis in UCINET. UCINET generate the principal eigenvector which the elements are taken as the ranking values for location nodes and presented in a log file. These ranking values produced by the benchmark system are termed as benchmark values of Dengue Hotspot Ranking (DHR) value, DHR_B. The location nodes are ranked based on the DHR_B and this ranking is the benchmark ranking of location nodes. Table 1 shows the benchmark ranking result and Table 2 the ranking result obtained in Bipartite Dengue Contact (BDC) network model ranking that was published earlier [3]. It showed that the benchmark ranking is the same as the BDC network model ranking.

Table 1. Benchmark Ranking

Location Node	$\mathbf{DHR_B}$
L4	1.000
L12	0.947
L17	0.936
L19	0.881
L3	0.000
L11	0.000
L16	0.000
L7	0.000
L6	0.000
L18	0.000
L13	0.000
L10	0.000
L2	0.000
L15	0.000
L1	0.000
L5	0.000
L8	0.000
L9	0.000
L14	0.000

Table 2. BDC Network Ranking

Location Node	$\mathbf{DHR_{BDC\_Net\_}}$
L4	1.000
L12	0.947
L17	0.937
L19	0.881
L3	0.00162
L11	0.00144
L16	0.00143
L7	7.49×10^{-7}
L6	4.64×10^{-7}
L18	4.19×10^{-7}
L13	2.89×10^{-7}
L10	3.3×10^{-116}
L2	3.1×10^{-116}
L15	2.9×10^{-116}
L1	0.000
L5	0.000
L8	0.000
L9	0.000
L14	0.000

The second step is to execute RMSE analysis which both DHR_B and DHR_{BDC_Net} are the inputs to the RMSE Generator. RMSE for location nodes is defined in equation 2 where N_{Loc} = the number of location node = 19 in this research and i is a natural number starting from 1.

$$RMSE_{Loc}\left(DHR_{B}\,, DHR_{BDC\_Net}\right) = \sqrt{\frac{1}{N_{Loc}}\left[\sum_{i=1}^{N_{Loc}}(\{DHR_{B}\}_{i} - \{DHR_{BDC\_Net}\}_{i})^{2}\right]} \quad (2)$$

Next, model validation is conducted to check the accuracy of the results obtained from the BDC network models [3]. Targeted and validated models are defined in previous work and the details of the targeted and validated models are explained in Table 3.

Table 3. Targeted and validated model details.

Group	BDC Network	Model	Epi Week	Number of Human Nodes	Number of Location nodes
1	1	Targeted model	28-29	8 dengue patients	19 locations
	2	Validated model	30-31	3 dengue patients	27 locations
2	3	Targeted model	32-33	12 dengue patients	78 locations
	4	Validated model	34-35	2 dengue patients	81 locations
3	5	Targeted model	36-37	10 dengue patients	98 locations
	6	Validated model	38-39	7 dengue patients	100 locations

Based on the first group in Table 3, targeted network is a network model for Epi Week (EW) 28 and 29 which consists of 8 dengue patients and the locations were visited by the dengue patients while the subsequent EW (EW 30-31) from the targeted model is used to build a validated model to compare the correlation between these location nodes in subsequent week. To validate it, the ranking result yielded by the targeted model should agree with the result obtained from the validated model.

Spearman Ranking Correlation Coefficient (SRCC) is used to measure the closeness of the ranking for the small size network [17-19]. The value of SRCC can be interpreted as Table 4. Based on Table 4, SRCC values range from -1 until +1. A positive correlation coefficient expresses a positive association between the ranking of two network models while a negative correlation coefficient expresses a negative association between the ranking. A correlation coefficient of 0 indicates that no association between the rankings exists at all. Regardless whether the value is positive or negative, the higher the value of SRCC, the stronger the association between the ranking of the two network models. In this study, the threshold SRCC value set at more than 0.70, indicating high correlation between the targeted result and observed data. Therefore, we can conclude that with SRCC > 0.70, the BDC network model is validated.

Table 4. SRCC Indicator [17]

SRCC coefficient	Indicator
± (0.00-0.19)	Very Weak
± (0.20-0.39)	Weak
± (0.40-0.59)	Moderate
± (0.60-0.79)	Strong
± (0.80-1.00)	Very Strong

3 Results and discussion

To verify the model, RMSE calculations on location nodes are tabulated in Table 5. The sum of squares difference obtained from Table 5 is substituted into equation 2. As a result, RMSE for location node is 0.0006419 is shown in equation 3, correct to four significant figures. The RMSE is much lesser than the threshold value of 0.05.

Table 5. RMSE Analysis of Location Nodes

Location Node	$\mathbf{DHR_B}$	$\mathbf{DHR_{BDC_{Net}}}$	$\mathbf{DHR_B - DHR_{BDC_{Net}}}$	$\mathbf{(DHR_B - DHR_{BDC_{Net}})^2}$
L1	0	0	0	0
L2	0	3.10×10^{-116}	-3.10×10^{-116}	9.61×10^{-232}
L3	0	0.00162	-0.00162	2.61×10^{-6}
L4	1	1	0	0
L5	0	0	0	0
L6	0	4.64×10^{-7}	-0.000000464	2.15×10^{-13}
L7	0	7.49×10^{-7}	-0.000000749	5.61×10^{-13}
L8	0	0	0	0
L9	0	0	0	0
L10	0	3.30×10^{-116}	-3.30×10^{-116}	1.09×10^{-231}
L11	0	0.00144	-0.00144	2.07×10^{-6}
L12	0.947	0.947	-2.77×10^{-5}	7.67×10^{-10}
L13	0	2.89×10^{-7}	-0.000000289	8.35×10^{-14}
L14	0	0	0	0
L15	0	2.90×10^{-116}	-2.90×10^{-116}	8.41×10^{-232}
L16	0	0.00143	-0.00143	2.04×10^{-6}
L17	0.936	0.937	-0.00105	1.10×10^{-6}
L18	0	4.19×10^{-116}	-0.000000419	1.76×10^{-13}
L19	0.882	0.881315	4.09×10^{-5}	1.68×10^{-9}
Sum of $(\mathrm{DHR_B} - \mathrm{DHR_{BDC_{Net}}})^2$:				7.83×10^{-6}

$$RMSE_{Loc}\left(DHR_B\ DHR_{BDC_{Net}}\right) = \sqrt{\tfrac{1}{13}[7.83 \times 10^{-6}]} = 0.0006419 \tag{3}$$

It is observed that the RMSE of location nodes (0.0006419) are much lesser than the threshold RMSE value of 0.05. Therefore, it can be concluded that the BDC network model formulated in this study is a verified model.

The BDC network model for group 1 of Table 3 has been validated in previous work [3]. Hence, this paper will consist the validation results for the second and third groups. In group 2, the targeted model built by using the data on EW 32 and 33 consist of 12 dengue patients and the patients visited 78 locations. From these 78 locations, there are 51 new incoming location nodes found and added into the network model. The database of location is extended from 19 location nodes (during the implementation of the first BDC network) to 27 location nodes (during the implementation of the second BDC network) and now there are 78 location nodes (during the implementation of the third BDC network). On the other hand, the validated model in group 2 utilized the data on EW 34 and 35 that consist of 2 human nodes and 81 location nodes. There are 3 new incoming location nodes added.

The Dengue Contact Strength (DCS) values for targeted and validated model in group 2 are calculated by using the procedures from previous study [3]. The locations' ranking results for targeted and validated model are tabulated in Table 6 and Table 7.

From both Table 6 and 7, there are four identical location nodes which include L53, 64, 69 and 77 in targeted and validated model. Thus, these four location nodes are used to calculate the SRCC value as depicted in Table 8. Equation 4 gives the formula used for the calculation of ρ_{Group2} where N_{Loc} =4 and a is a natural start from 1. Table 8 presents the calculation of various terms in equation 4.

$$\rho = 1 - \frac{6\sum_{a=1}^{N_{Loc}}[\{d\}_a]^2}{N_{Loc}\left({N_{Loc}}^2 - 1\right)}$$

$$\text{where } \{d\}_a = \left\{RankDHR_{Targeted}\right\}_a - \{RankDHR_{Validated}\}_a \tag{4}$$

Table 6. Location Node Ranking of the Targeted Model in Group 2

Rank	Location Node	DHR
1	L63	1
2	L78	0.8881
3	L55	0.8379
4	L73	0.8296
5	L62	0.0005
6	L52	0.0003
7	L33	0.0003
8	L77	0.0003
9	L71	0.0002
10	L59	0.0002
11	L67	9.4×10^{-5}
12	L57	8.4×10^{-5}
13	L66	2.5×10^{-110}
14	L74	2.2×10^{-110}
15	L56	2.2×10^{-110}
16	L69	4.0×10^{-114}
17	L60	3.8×10^{-114}
18	L76	3.2×10^{-114}
19	L58	4.9×10^{-123}
20	L68	4.2×10^{-123}
21	L75	3.6×10^{-123}
22	L23	1.4×10^{-124}
23	L47	1.2×10^{-124}
24	L53	1.1×10^{-124}
25	L54	3.7×10^{-135}
26	L72	3.6×10^{-135}
27	L19	3.5×10^{-135}
28	L61	2.6×10^{-203}
29	L70	2.5×10^{-203}
30	L64	1.4×10^{-299}

Table 7. Location Node Ranking of the Validated Model in Group 2

Rank	Location Node	DHR
1	L80	1.0000
2	L77	0.9784
3	L69	0.8595
4	L81	0.8556
5	L64	0.6889
6	L53	2.2×10^{-6}
7	L79	1.8×10^{-6}

$$\rho = 1 - \frac{6\sum_{a=1}^{N_{Loc}}[\{d\}_a]^2}{N_{Loc}\left(N_{Loc}{}^2 - 1\right)}$$

$$\text{where } \{d\}_a = \left\{RankDHR_{Targeted}\right\}_a - \{RankDHR_{Validated}\}_a \tag{4}$$

Table 8. Calculation of SRCC for Location Nodes between Targeted and Validated Model in BDC Network Model 2

Location Node	Rank $\mathrm{DHR_{Targeted}}$	Rank $\mathrm{DHR_{Validated}}$	Rank $\mathrm{DHR_{Targeted}}$ - Rank $\mathrm{DHR_{Validted}}$, d	d^2
L53	3	4	-1	1
L64	4	3	1	1
L69	2	2	0	0
L77	1	1	0	0
			Sum of d^2:	2
			ρ_{Group2}:	0.8

For group 3, the targeted model built by using the data on EW 36 and 37 which consists of 10 human nodes and 98 location nodes while the validated model utilized the data on EW 38 and 39 that consist of 7 human nodes and 100 location nodes. Similarly, the locations' ranking for targeted and validated models in group 3 of Table 3, are calculated and the SRCC value is computed. Table 9 presents the SRCC values of all three groups of network model. The SRCC values of these three groups are higher than the threshold value 0.70. Thus, the BDC network models are validated.

Combining the results produced from the verification and validation analysts of the results, it can be concluded that the BDC network model is now verified and validated. Researcher obtained a strong ranking similarity to verify and validate bipartite network model by using the similar approach [20].

Table 9. BDC Network Model Validation

Group	BDC Network	SRCC, ρ
1	1 2	1.0000
2	3 4	0.8000
3	5 6	0.8424

4 Conclusion

In this study, the bipartite dengue contact (BDC) network has been verified and validated. To conclude, the bipartite network modelling approach has been successfully formulated and the ranked locations are believed to be applicable to help public health authorities to prioritise the locations for vector control. Eradication of dengue in the risk areas can help in the reduction of the spread of dengue disease.

Acknowledgements

The authors thank Universiti Malaysia Sarawak for the support in carrying out this research under the grant numbered F08/SpFRGS/1601/2017. Our heartfelt thanks also go to Sarawak State Health Department and Sarawak Meteorological Department for providing the research data.

References

1. Hassarangsee, S., Tripathi, N. K., & Souris, M. (2015). Spatial pattern detection of tuberculosis: a case study of Si Sa Ket Province, Thailand. *International journal of environmental research and public health*, *12*(12), 16005-16018.
2. Rueda, L. M., Patel, K. J., Axtell, R. C., & Stinner, R. E. (1990). Temperature-dependent development and survival rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae). *Journal of medical entomology*, *27*(5), 892-898.
3. Kok W.C., Labadin J., Perera D. (2018) Modeling Dengue Hotspot with Bipartite Network Approach. In: Alfred R., Iida H., Ag. Ibrahim A., Lim Y. (eds) Computational Science and Technology. ICCST 2017. Lecture Notes in Electrical Engineering, vol 488. Springer, Singapore.
4. Carrington, L. B., Armijos, M. V., Lambrechts, L., & Scott, T. W. (2013). Fluctuations at a low mean temperature accelerate dengue virus transmission by Aedes aegypti. *PLoS neglected tropical diseases*, *7*(4), e2190.
5. Tun-Lin, W., Burkot, T. R., & Kay, B. H. (2000). Effects of temperature and larval diet on development rates and survival of the dengue vector Aedes aegypti in north Queensland, Australia. *Medical and veterinary entomology*, *14*(1), 31-37.
6. Focks, D. A., Patz, J. A., Martens, W. J., & Jetten, T. H. (1998). Dengue fever epidemic potential as projected by general circulation models of global climate change. *Environmental health perspectives*, *106*(3), 147.
7. Phaijoo, G. R., & Gurung, D. B. (2015). Mathematical Study of Biting Rates of Mosquitoes in Transmission of Dengue Disease. *Journal of Science. Engineering and Technology*, *11*, 25-33.
8. Sylvestre, G., Gandini, M., & Maciel-de-Freitas, R. (2013). Age-dependent effects of oral infection with dengue virus on Aedes aegypti (Diptera: Culicidae) feeding behavior, survival, oviposition success and fecundity. *PloS one*, *8*(3), e59933.
9. Scott, T. W., Amerasinghe, P. H., Morrison, A. C., Lorenz, L. H., Clark, G. G., Strickman, D., ... & Edman, J. D. (2000). Longitudinal studies of Aedes aegypti (Diptera: Culicidae)

in Thailand and Puerto Rico: blood feeding frequency. *Journal of medical entomology*, *37*(1), 89-101.

10. Davis, P. K. (1992). Generalizing concepts and methods of verification, validation, and accreditation (VV&A) for military simulations (No. RAND/R-4249-ACQ). RAND CORP SANTA MONICA CA.
11. Cook, D. A., & Skinner, J. M. (2005). How to perform credible verification, validation, and accreditation for modeling and simulation. *The Journal of Defense Software Engineering*.
12. Liew, C. Y. (2016). *Bipartite Network Modeling of Habitat Suitability*. (Unpublished doctoral dissertation). Universiti Malaysia Sarawak, (UNIMAS).
13. Albright, J. J., & Park, H. M. (2009). Confirmatory factor analysis using amos, LISREL, Mplus, SAS/STAT CALIS.
14. Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230-258.
15. Eze, M. O. (2013). *Web Algorithm search engine based network modelling of Malaria Transmission.* (Doctoral dissertation) Universiti Malaysia Sarawak (UNIMAS).
16. Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for social network analysis.
17. Bouzerdoum, A., Havstad, A., & Beghdadi, A. (2004). Image quality assessment using a neural network approach. In *Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on* (pp. 330-333). IEEE.
18. Lim, W. K., Wang, K., Lefebvre, C., & Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, *23*(13), i282-i288.
19. Tetko, I. V., & Tanchuk, V. Y. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of chemical information and computer sciences*, *42*(5), 1136-1145.
20. Liew, C., & Labadin, J. (2017). Applying Bipartite Network Approach to Scarce Data: Validation of the Habitat Suitability Model of a Marine Mammal Species. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *9*(3-11), 13-16.

Comparison of Classification Algorithms on ICMPv6-Based DDoS Attacks Detection

Omar E. Elejla[1], Bahari Belaton[1], Mohammed Anbar[2], Basim Alabsi[2] and Ahmed K. Al-Ani[2]

[1] School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia,
[2] National Advanced IPv6 Centre (NAv6), Universiti Sains Malaysia, Penang, Malaysia
oeoe14_com063@student.usm.my, bahari@usm.my, anbar@nav6.usm.my,
baalabsi77@gmail.com, ahmedKhallel91@nav6.usm.my

Abstract. Computer networks are aimed to be secured from any potential attacks. Intrusion Detection systems (IDS) are a popular software to detect any possible attacks. Among the mechanisms that are used to build accurate IDSs, classification algorithms are extensively used due to their efficiency and auto-learning ability. This paper aims to evaluate classification algorithms for detecting the dangerous and popular IPv6 attacks which are ICMPv6-based DDoS attacks. A comparison between five classification algorithms namely Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Neural Networks (NN) were conducted. The comparison was conducted using a publicly available flow-based dataset. The experimental results showed that classifiers have detected most of the included attacks with a range from 73%-85% for the true positive rate. Moreover, KNN classification algorithm has been the fastest algorithm (0.12 seconds) with the best detection accuracy (85.7%) and less false alarms (0.171). However, SVM achieved the lowest detection accuracy (73%) while NN was the slowest algorithm in training the detection model (323 seconds).

Keywords: Intrusion Detection systems, IPv6, ICMPv6, Attacks Detection, Decision Tree, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, Neural Networks

1 Introduction

Internet Protocol version six (IPv6) has been proposed with enhanced security and communication features to eventually replace Internet Protocol version four (IPv4). However, it has been attacked by different types of attacks whereas the number of networks that experienced IPv6 attacks is being increased to be 13% in 2016 while it was 9% in 2015 [1]. IPv6 suffers from different types of attacks which are either similar to the IPv4 attacks or new attacks that are appeared recently with IPv6 new features [2]. These attacks expose IPv6 adoption and would be the reason for slowing down its deployment in the existing networks if they are not correctly addressed [3]. According to experiment by Ard [4], Denial of Service (DoS) attacks (including Dis-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_34

tributed DoS attacks) are the most performed attacks against IPv6 network among all IPv6 vulnerabilities classes.

ICMPv6 is the core part of IPv6, responsible for the core communications between the nodes and with the routers. ICMPv6 is a mandatory protocol to be implemented in the networks in order to use IPv6 for the communication. The mandatory availability of ICMPv6 protocol made it a preferred medium for attackers to be used in attacking IPv6 networks. Due to the popularity of IPv6 Distributed DoS (DDoS) attacks and the importance of ICMPv6 protocol, DDoS attacks that are targeting ICMPv6 (ICMPv6-based DDoS) attacks are security priority to be addressed [5].

One of the possible ways to detect IPv6 attacks is to monitor the network traffic looking for any illegal traffics or behaviors which are called intrusions. Intrusion Detection System (IDS) is the responsible software for automating these tasks for a network or node that it is installed on [6]. IDSs are installed on an edge point of the network to monitor the whole passing traffic and alert the administrator of any suspicious behavior (activity). The IDSs are classified based on the followed detection mechanism into signature-based IDSs (SIDSs) and anomaly-based IDSs (AIDSs). SIDSs depend on a pattern for each attack that indicates its existence. Second, AIDSs define a profile of the allowed behaviors in the network and any deviation is suspicious behavior. Unlike SIDS, AIDSs are unable to detect “zero-day” attacks (their signatures are not recorded in the SIDSs’ database). Therefore, AIDS is a good choice to detect ICMPv6-based DDoS attacks as it recognizes the behaviors of the attacks and provides the ability to detect unknown attack [2, 7].

AIDSs work based on the assumption that intrusions generate abnormal activities that indicate their existence, thus they try to differentiate between normal and abnormal behaviors. Therefore, AIDS is considered as a classification problem that aims to train a model to learn how to differentiate between normal and malicious traffic. Classification algorithms are considered as the most efficient techniques that show impressive and reliable results in building these models[8]. Moreover, these algorithms have the ability to automate the process of building the detection models and to diminish human effort required to build these model [9]. Therefore, AIDSs extensively use classification algorithms in their detection which proved their ability to accurately detect attacks on many computer networks [10].

IDSs are evaluated and compared prior to their installation using labeled datasets that include the possible scenarios of the targeted attack scenarios. IPv6 security suffers from lack availability of benchmark datasets. The existing IPv6 IDSs were applied to self-generated datasets that their comprehensiveness and completeness are not guaranteed. Moreover, the existing benchmarked IPv4 datasets such as DARPA [11] and NSL_KDD [12] cannot be used for modeling IPv6 IDSs due to the specification differences between the two protocols.

In this paper, five classification algorithms are compared and evaluated in the ability to detect the ICMPv6-based DDoS attacks. The compared algorithms are Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Neural Networks (NN). These algorithms have been applied and compared in detecting IPv4 attacks and showed impressive performances while they have not been compared to detect IPv6 attacks. Two datasets of ICMPv6-based DDoS attacks have been created and made online available for others researcher by Elejla, Anbar [13]. Based on the literature, these datasets are the most suitable choice for conducting the experimental comparison. The comparison used Elejla, Anbar [13] datasets which include the ICMPv6-based DDoS attacks in flow-based representation. To the best of the authors' knowledge, this comparison is the first comparison of these classification algorithms that aims to practically evaluate them in detecting IPv6 attacks (ICMPv6-based DDoS attack).

The rest of the paper is organized as follows; Section 2 presents a review of the chosen classification algorithms. Section 3 discusses the usability of the existing datasets as well as the characteristics of the used dataset, the evaluation metrics, and the experimental results. Section 4 concludes the findings of the paper.

2 Classification Algorithms

Classification is the process of training the classifier in labeled traffic to learn the differences between the included classes and come up with a detection model [14]. This model is tested using another network traffic dataset to measure its prediction ability of the traffic. Based on the testing results, the detection model is integrated into real networks to detect new traffic of the attacks that it trained on [15, 16]. This section presents a brief overview of the used classification algorithms in the comparison which are Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Neural Networks (NN). The classifiers are chosen as they are common and available in WEKA platform [17], thus they should not be implemented again. Moreover, they cover different commonly used classification methods, i.e., decision tree, uncertainty modeling, example-based, machine learning and probabilistic models.

2.1 Decision Tree (DT)

DT is one of the most popular classification algorithms that has been used to solve several problems in different areas. It has an advantage among other algorithms that it is easy to interrupt its structure, robust to curse of dimensionality and robust to data noise. Moreover, it does not require prior knowledge of the traffic, unlike other algorithms which might require traffic distribution model and parameters [18]. To the best of the author's knowledge, DT tress has not been applied to detect IPv6 attacks thus it will be applied for the first time in this paper

2.2 Support Vector Machine (SVM)

SVM is a supervised learning algorithm proposed by Corinna Cortes and Vapnik (1995), and it became one of the most popular approaches in the classification learning area. It has been applied to various applications such as pattern recognition, text categorization, image classification, etc. [19-21]. SVM's basic concept depends on structural risk minimization. It uses a nonlinear mapping in order to transfer the input training pattern into a high dimensional feature space where the optimal separating hyperplane can be found. SVM has been used by Zulkiflee, Haniza [22] and Anbar, Abdullah [23] to detect flooding attacks of ICMPv6 RA messages using 5 and 9 packet-based features respectively. However, the used packet-based representation and the features are unsuitable for such attacks detection as shown in [2, 24].

2.3 Naïve Bayes (NB)

NB is one of the probabilistic classification algorithms has been proposed to take advantages from the structural relation (dependencies) between the targeted problem's variable, especially for uncertainty domain problems. The idea of NB depends on calculating the probability of one class label when a particular behavior (event) exists [14]. NB is a fast classification algorithm that used a graphical modeling mechanism to represent both causal and probabilistic relationships (interdependencies) between the included events, thus it has the ability to model problems that need prior knowledge of the problems [15]. NB has been enhanced and applied to detect IPv6 covert channel attacks by Salih, Ma [25]. Covert channels attacks are performed by using unused flags or bits in the packets to send the malicious data toward the victims to avoid the security mechanisms. However, Salih, Ma [25]'s work has been criticized as it did not address other kinds of IPv6 attacks such as DDoS attacks. In addition, it depends on a packet-based representation of the traffic and non-qualified features such as IPv6 source address [2].

2.4 K-Nearest Neighbors (KNN)

KNN is a similarity-based classification algorithm, works by predicting the class label based on the similarity between its given records (examples) in the training traffic. Based on the calculated distances between the given points on the input traffic, it determines the K-nearest neighbors for the unlabeled traffic. K value (the number of the nearest neighbors) is an important factor that affects the training time and the performance of the algorithm. It is one of the simplest algorithms that does not build a training model. It works by an "on-line" mechanism that searches the nearest neighbor for the unlabeled record to determine its class [26]. To the best of the author's knowledge, KNN has not been applied to detect IPv6 attack thus it will be applied for the first time in this paper

2.5 Neural Networks (NN)

NN is inspired by the neurons of human brains to predict and classify data. It is designed with interconnected nodes with a weight for each connection between them. Each node receives the input from its interconnected nodes to compute the output function and send it to the next interconnected nodes. A number of nodes and layers are tuned parameters can be input by the users. Multilayer Perceptron (MLP) is the most used architecture of NN that achieve accurate results in different applications [14, 26]. The backpropagation learning algorithm is combined with the MLP to train it and that is called as Backpropagation Neural Networks (BPNN). BPNN has been applied by Saad, Anbar [27] to detect ICMPv6 echo request flooding attacks. However, this works has been criticized by Elejla, Belaton [2] in terms of the used packet-based representation and features.

3 Experimental result and discussion

This section aims to apply the classification algorithms to evaluate their effectiveness in detecting the ICMPv6-based DDoS attacks. Section 3.1 describes the details and characteristic of the used dataset. Section 3.2 highlights the evaluation metrics that have been used in the experiments. Section 3.3 presents and discusses the experimental results of applying the classification algorithms.

3.1 Dataset

As mentioned, the many IPv4 datasets cannot be used in this paper because they are free of IPv6 traffic which is the aim of this research. Moreover, several IPv6 datasets have been created in different IPv6 researches to achieve their authors' requirements. However, these datasets have several drawbacks mentioned in [13, 28] that limit their use. These drawbacks are as follow; MAWILab dataset [29] does not include any attack traffic, [30] dataset (Ark IPv6 topology dataset) is an unlabeled dataset. Other datasets have been proposed by Zulkiflee, Haniza [22], SAAD, MANICKAM [31] and Najjar and Kadhum [32] are inappropriate to be used as they are unavailable for us to use as well as they do not include the targeted attacks (ICMPv6-based DDoS attacks). On the other side, recently Elejla, Anbar [13] have proposed flow-based datasets which are labeled datasets, available online on [33], covered diverse scenarios of ICMPv6-based DDoS attacks. Moreover, the dataset fulfilled the requirements of the good dataset which are realistic traffic, diverse scenarios, completely and correctly labeled, sufficient and balanced size and representative features. All features were normalized to numeric datatype by their author to meet all the possible classifiers.

Classification algorithms are evaluated using labeled datasets to determine their ability to model the attacks as well as predict new unlabeled attacks. To apply, evaluate and compare the chosen classification algorithm, Elejla, Anbar [13] dataset has been used due to their online availability, containing the targeted attacks and labeled traffic. The dataset contains 101,088 records (flows) with 49,187 attack records and 51,901 normal records. Moreover, the dataset has been preprocessed (balanced and normalized) to be ready for applying the classifiers. Table 1 shows the specifications of the used flow-based dataset.

Table 1. The Flow-based Dataset Specifications

Source	USM university, School of computer sciences laboratory network
Number of features	11 features
Representation	Flow-based representation
Number of attack flows	49,187 attack flows
Number of normal flows	51,901 normal flows
Attacking tools	The Hacker Choice's IPv6 (THC-IPv6) [34] and SI6 [35]
Number of attacks scenarios	22 attacks scenarios

The dataset includes several types of ICMPv6-based DDoS attacks that were performed in a real network. The dataset traffic has been represented using a flow-based representation with 11 flow-based features. Elejla, Anbar [13] composed a flow by combing the packets that shared IPv6 source and destination, port source and destination, and the protocol in one record (flow). The 11 features have been reasonably selected with logical justifications of their direct relation to the attacks detection. Table 2 shows the features that represent each flow in the dataset with their description.

Table 2. The Flow-based Features with their Description

#	Feature Name	Description
1	ICMPv6type	ICMPv6 type of the flow's packets.
2	PacketsNumber	Number of transferred packets within the flow.
3	TransferredBytes	Number of bytes sent from the source to the destination.
4	Duration	Time Length of the flow
5	Ratio	Ratio of bytes transferring during the flow duration.
6	Length_STD	Variation in the Length of the flow's packets.
7	FlowLabel_STD	Variation in the Flow Label of the flow's packets.
8	HopLimit_STD	Variation in the Hop Limit of the flow's packets.
9	TrafficClass_STD	Variation in the Traffic Class of the flow's packets.
10	NextHeader_STD	Variation in the Next Header of the flow's packets.
11	PayloadLength_STD	Variation in the Payload Length of the flow's packets.

3.2 Evaluation Metrics

The ability of the classification algorithms in detecting ICMPv6-based DDoS attacks is measured in term of several metrics. The used evaluation metrics are Classification Accuracy (CA), True Positive Rate (TPR) or Recall, False Positive Rate (FPR), Precision, F-Measure, ROC Area and training time. These metrics are calculated using few parameters that have been described and shown in Table 3.

Table 3. Description of the Evaluation Metrics Parameters

Evaluation Metric	Description
True Positive (TP)	Number of samples predicted as attack that are actually attack
False Positives (FP)	Number of samples predicted as attacks that are actually normal
True Negatives (TN)	Number of samples predicted as normal that are actually normal
False Negatives (FN)	Number of samples predicted as normal that are actually attack

1. **Classification Accuracy (CA)** is the percentage of the correctly classified samples from the total number of samples. It can be calculated using Equation 1.

$$CA = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \tag{1}$$

2. **True Positive Rate (TPR) or Recall** is the percentage of the correctly detected attack samples from the total number of attacks samples. It can be calculated using Equation 2.

$$TPR = \frac{TP}{FN+TP} * 100\% \tag{2}$$

3. **False Positive Rate (FPR)** is the percentage of the normal samples that are misclassified as attacks from the total number of normal samples. It can be calculated using Equation 3.

$$FPR = \frac{FP}{TN+FP} \tag{3}$$

4. **Precision** is the percentage of the correctly detected attack samples from the total number of samples that are classified as attacks. It can be calculated using Equation 4.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

5. **F-Measure** is defined as the weighted harmonic mean of the precision and recall of the values. F -measure has value ranges from 0 to 1; 1 means that decision of attacks is accurate. It can be calculated using Equation 5.

$$F\text{-}Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

6. **Training time** is the time needed for the classifier to be trained on the dataset to build the detection model. It reflects the speed of the classifier which is a key factor especially when retraining of the model is needed to update it.

3.3 Results and Discussion

This section presents and discusses the experimental results of applying the classification algorithms to the aforementioned flow-based dataset (Section 3.1) based on the mentioned evaluation metrics (Section 3.2). The experiments were conducted using WEKA 8.3 platform on a computer with the hardware specifications that are mentioned in Table 4.

Table 4. Hardware Specifications

CPU	Intel(R) Core(TM) i7-2670QM CPU 2.20GHz @ 2.20GHz
Memory	6.00GB
Operating system	Windows 7 (64bit)

Table 5 shows the evaluation metrics results of applying the classification algorithms to the dataset. The classifiers are applied using 10 folds Cross-validation testing approach which divides the dataset into two parts; 90% for training the model and 10% for testing the model's prediction ability. The classification algorithms have been applied with their default parameters that are available in WEKA without any parameter tuning or optimization. Each of the classification algorithms has been applied for 10 times and the average of each metric has been calculated and presented in Table 5.

Table 5. The Experimental Results of Applying the Classification Algorithms to the Dataset

Classifier Name	WEKA Classifier Name	CA	TPR	FPR	Precision	F-Measure	Training Time (seconds)
DT	J48	85.7	0.857	0.171	0.885	0.852	9.15
SVM	SMO	73.5	0.735	0.292	0.746	0.727	50.5
NB	Naive Bayes	74.5	0.745	0.300	0.805	0.724	0.63
KNN	IBK	85.7	0.857	0.171	0.885	0.852	0.12
NN	Multilayer Perceptron	83.2	0.832	0.197	0.859	0.826	323.29

As shown in Table 5, the classification algorithms have achieved different values of the evaluation metric. However, some of them have achieved better values compared to the others. First, DT and KNN algorithms have hit the best rates in term of most of the evaluation metrics compared to the other algorithms. However, DT needed longer time to train its model compared to KNN. Second, NN has achieved lower values in term of CA, TPR, FPR, Precision, and F-measure compared to KNN and DT. However, it needed the longest time to train its model compared to other classification algorithms. Lastly, SVM and NB achieved almost the same evaluation metrics expect the training time whereas SVM completed training in a longer time than NB. SVM and NB have been the worst in terms of CA, TPR, FPR, Precision and F measure among other algorithms.

In a nutshell, in term of the detection accuracy and low false alarm, KNN and DT have outperformed other algorithms. KNN and DT have shown that they are able to detect most of the available ICMPv6-based DDoS attacks that might be because KNN and DT benefit from the overfitting existence in the dataset. Furthermore, KNN has beaten other in term of training time too. However, SVM has given the lowest detection accuracy with second longer training time as SVM does not benefit from overfitting existence. In addition, NB has given the highest false positive rates with a detection ability close to the SVM ability. NN has given a moderate detection ability and false positive rate but has taken long training time. These long training times might be due to the lengthy processes needed by the NN to choose the best parameters to build its inner neural.

4 Conclusion

Due to the importance of the classification process in building a reliable IDS as well as the real danger of ICMPv6-based DDoS attack, classification algorithms have been applied for detecting them. Several classification algorithms that follow different classification mechanisms have been applied to detect ICMPv6-based DDoS attacks.

The algorithms show that the used dataset with its representation and features are suitable for DDoS attacks detection as most of the attacks have been detected by the algorithms with varying detection abilities. The algorithms have been compared in terms of the classification accuracy, true positive rate, false positive rate, precision, f-measure and training time. KNN has achieved the best detection ability with low false positive rates in addition to fast training time compared to the rest of the classification algorithms.

The experimental results are comparatively acceptable as the algorithms were able to detect more than 73.5% of the ICMPv6-based attacks. Moreover, the total detection accuracies were larger than 73.5% with 30% false alarms in the worst case. However, these results did not reach a reliable level that qualifies the build models to be implemented in real IDSs. More efforts need to be made to improve these results by either optimizing the classification algorithms or tune the classifiers' parameters. In addition, more features might be added to the dataset to increase the detection ability of the classifier to recognize the attacks.

As an extension of this paper, different paths are potential to be followed to improve the proposed experimental comparison. First, there is a need to find the best parameters for each classification algorithm in order to improve their detection performances. Finding these parameters might be selected experimentally or by applying parameter tuning algorithms. Second, choose other classification algorithms to be included in this comparison to discover their efficiency in detecting the attacks. Last, enrich the dataset with extra features that can help the classifiers to differentiate between the attack and normal records. These features might be determined based on studying the attacks domain-knowledge or adopting from other similar attacks features.

3 Acknowledgment

The authors would like to thank the School of Computer Science, Universiti Sains Malaysia (USM) for providing the facilities and support. This research was supported by the USM RUI Grant 1001/PKOMP/8014018.

4 References

1. Anstee, D., et al., Worldwide Infrastructure Security Report. 2017, ARBOR Network
2. Elejla, O.E., et al., Intrusion Detection Systems of ICMPv6-based DDoS attacks. Neural Computing and Applications, 2016: p. 1-12.
3. Caicedo, C.E. and J. Joshi, Security issues in ipv6 networks. International Telecommunications Research and Education Association (ITERA), 2008.
4. Ard, J.B., Internet protocol version six (ipv6) at uc davis: traffic analysis with a security perspective. 2012, University of California, Davis.
5. Elejla, O.E., M. Anbar, and B. Belaton, ICMPv6-based DoS and DDoS attacks and defense mechanisms. IETE Technical Review, 2017. 34(4): p. 390-407.
6. Scarfone, K. and P. Mell, Guide to intrusion detection and prevention systems (idps). NIST special publication, 2007. 800(2007): p. 94.

7. Shon, T. and J. Moon, A hybrid machine learning approach to network anomaly detection. Information Sciences, 2007. 177(18): p. 3799-3821.
8. Elejla, O.E., et al., Flow-Based IDS for ICMPv6-Based DDoS Attacks Detection. Arabian Journal for Science and Engineering, 2018.
9. Shamshirband, S., et al., An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique. Engineering Applications of Artificial Intelligence, 2013. 26(9): p. 2105-2127.
10. Anbar, M., et al. Comparative performance analysis of classification algorithms for intrusion detection system. in Privacy, Security and Trust (PST), 2016 14th Annual Conference on. 2016. IEEE.
11. Lippmann, R., et al., The 1999 DARPA off-line intrusion detection evaluation. Computer Networks, 2000. 34(4): p. 579-595.
12. Stolfo, S.J., et al. Cost-based modeling for fraud and intrusion detection: results from the JAM project. in DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings. 2000.
13. Elejla, O.E., et al., Labeled flow-based dataset of ICMPv6-based DDoS attacks. Neural Computing and Applications, 2018.
14. Agrawal, S. and J. Agrawal, Survey on anomaly detection using data mining techniques. Procedia Computer Science, 2015. 60: p. 708-713.
15. Patcha, A. and J.-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks, 2007. 51(12): p. 3448-3470.
16. Muniyandi, A.P., R. Rajeswari, and R. Rajaram, Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. Procedia Engineering, 2012. 30: p. 174-182.
17. Witten, I.H., et al., Data Mining: Practical machine learning tools and techniques. 2016: Morgan Kaufmann.
18. Hodge, V. and J. Austin, A survey of outlier detection methodologies. Artificial intelligence review, 2004. 22(2): p. 85-126.
19. Burges, C.J., A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998. 2(2): p. 121-167.
20. Joachims, T., Text categorization with support vector machines: Learning with many relevant features. 1998: Springer.
21. Chapelle, O., P. Haffner, and V.N. Vapnik, Support vector machines for histogram-based image classification. Neural Networks, IEEE Transactions on, 1999. 10(5): p. 1055-1064.
22. Zulkiflee, M., et al., A Framework of IPv6 Network Attack Dataset Construction by Using Testbed Environment. International Review on Computers and Software (IRECOS), 2014. 9(8).
23. Anbar, M., et al., A Machine Learning Approach to Detect Router Advertisement Flooding Attacks in Next-Generation IPv6 Networks. Cognitive Computation, 2017: p. 1-14.
24. Elejla, O.E., et al. A New Set of Features for Detecting Router Advertisement Flooding Attacks. in Information and Communication Technology (PICICT), 2017 Palestinian International Conference on. 2017. IEEE.
25. Salih, A., X. Ma, and E. Peytchev, Detection and Classification of Covert Channels in IPv6 Using Enhanced Machine Learning. 2015.
26. Tsai, C.-F., et al., Intrusion detection by machine learning: A review. Expert Systems with Applications, 2009. 36(10): p. 11994-12000.
27. Saad, R.M., et al., An intelligent icmpv6 ddos flooding-attack detection framework (v6iids) using back-propagation neural network. IETE Technical Review, 2016. 33(3): p. 244-255.

28. Elejla, O.E., et al., A Reference Dataset for ICMPv6 Flooding Attacks. Journal of Engineering and Applied Sciences, 2016. 100(3): p. 476-481.
29. Fontugne, R., et al., MAWILab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking, in Proceedings of the 6th International COnference. 2010, ACM: Philadelphia, Pennsylvania. p. 1-12.
30. CAIDA. The cooperative association for internet data analysis. 2014 2014 [cited 2017 28/02/2017]; Available from: https://www.caida.org/data/active/ipv6_allpref_topology_dataset.xml.
31. SAAD, R., et al., DESIGN & DEPLOYMENT OF TESTBED BASED ON ICMPv6 FLOODING ATTACK. Journal of Theoretical & Applied Information Technology, 2014. 64(3).
32. Najjar, F. and M.M. Kadhum. Reliable Behavioral Dataset for IPv6 Neighbor Discovery Protocol Investigation. in IT Convergence and Security (ICITCS), 2015 5th International Conference on. 2015. IEEE.
33. Elejla, O.E., M. Anbar, and B. Belaton. Flow-based Datasets 2016 [cited 2016; Available from: https://sites.google.com/site/flowbaseddatasets/.
34. Heuse, M. THC IPv6 attack tool kit. 2013 [cited 2015; Available from: http://www.aldeid.com/wiki/THC-IPv6-Attack-Toolkit.
35. Gont, F. Si6 networks' ipv6 toolkit. 2012 [cited 2015; Available from: http://www.si6networks.com.

Feedforward plus Feedback Control Scheme and Computational Optimization Analysis for Integrating Process

I.M.Chew[1], F.Wong[2], A.Bono [2], J.Nandong [1], and K.I.Wong [1]

[1] Curtin University Malaysia, Sarawak, Malaysia
[2] Universiti Malaysia Sabah, Sabah, Malaysia
chewim@curtin.edu.my, farrah@ums.edu.my, bono@ums.edu.my

Abstract. Integrating process is applied to many industries however there are very few research done on it. Determining PID settings for the closed-loop control of integrating process is a challenging task due to its inherent characteristic, which is only stable at one equilibrium operating point. This paper highlighted First Order plus Dead Time in representing process and disturbance model. Improvement of relative performance for transient and steady state response is achieved by using feedforward plus feedback control scheme. Moreover, computational optimization analysis was presented for developing a systematic way to design PID controller for the optimal performance of both servo and regulatory control problems. Performance of the controlled process were then compared in term of graphs, performance index and performance indicator. It is proven and concluded that designed PID controller settings by using computational optimization analysis eventually gives the best performance compared to other tuning methods for a Pumped-tank function of LOOP-PRO simulation software.

Keywords: Integrating process, Feedforward plus feedback control, Genetic Algorithm.

1 Introduction

1.1 Integrating Process and Feedforward plus Feedback Control Scheme.

Level control is a common controlled process in many industries particular for regulating water level of tank. To determine the controller settings of a closed-loop integrating process is surprisingly challenging due to inherent characteristic of integrating process that is only stable in an open loop configuration at its equilibrium operating point, where the total inflow rate is equal to total outflow rate of a tank [1,2]. The flow rate into the tank varies with time but the flow rate out is constantly regulated by a pump. If the inflow rate to the tank is not equal to the amount of outflow rate, the level of composition will continuous to vary and eventually the tank either is completely empties or overflows, unless one of the flow rates is immediately corrected.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_35

A typical feedback control loop can work well for servo control problem but not regulatory control problem due to the control action is only given after an external disturbance signal is received. It causes sluggish performance to steady state response. To overcome it, the feedforward control scheme is another alternative that can be proposed to the control system. Feedforward approach applies an additional sensor to measure a disturbance directly and integrated with process model for immediate feedforward control actions before the disturbance begins to change the process variable [4-7]. Block diagram of an embedded feedforward algorithm to the feedback control loop is illustrated in Fig. 1.

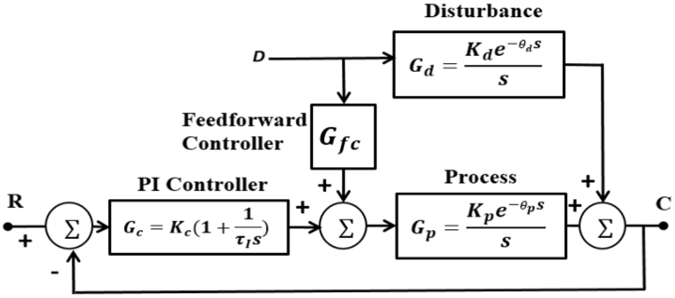

Disturbance
D
$G_d = \frac{K_d e^{-\theta_d s}}{s}$
Feedforward Controller
G_{fc}
PI Controller
$G_c = K_c(1 + \frac{1}{\tau_I s})$
Process
$G_p = \frac{K_p e^{-\theta_p s}}{s}$
R
C
+
-
+
+
+
+
Σ
Σ
Σ

Fig. 1. Block diagram of feedforward plus feedback control system.

Where
G_c = PID Controller
G_d = Disturbance Model
G_p = Process Model
G_{fc}=Feedforward Controller
K_d =Disturbance Gain
K_p =Process Gain
K_c= Proportional Gain
τ_I = Integral Time Constant
θ_p = Process Deadtime
θ_d = Disturbance Deadtime

In Fig.1, respective G_p and G_d is First Order plus Dead Time Integrating process, *FOPDT-integrating* model, whereby G_c is the controller to perform control action. The improvement is obtained through applying a feedforward controller where the tuning ratio is determined through dividing G_p with G_d.

The applied controller is Proportional-Integral-Derivative, *PID* controller. There are two *PID* optimal tunings in regards to both servo and regulatory control problems. The *PID* settings are subjected to the performed objective function however it confuses the operators of which setting should be applied. This paper solve the complication by proposing a tuning approach to obtain trade-off *PID* tunings that give the best performance for both servo and regulatory control. It can be achieved through computational optimization analysis. Relative performance of the proposed feedforward plus feedback control schemes for the integrating process were studied through Pumped Tank function of LOOP-PRO software, which is famously used in process control training [8].

1.2 Computational Optimization Analysis uses Genetic Algorithm.

Computational optimization analysis is widely used in analysing complex algorithm and optimization problems. Among them, Genetic Algorithm, *GA* is global searching technique uses genetic-based mechanism is iteratively repeated analysis for finding varies basin of attractions and eventually ended after meeting required tolerance that is denoted the optimum result of the problem [16]. It's overcome the findings of stuck in local minima. This great capability of *GA* is applied in designing *PID* controller settings as well as solving optimization problems. *GA* operates through five significant steps: initial population, fitness function, Selection, Crossover and Mutation [20]. Moreover, several settings were applied in GA analysis. The Scattered function of crossover is selected with ratio of 0.8, which is used to analysis optimum PI tunings whereby mutation is set at Constraint dependent.

From the literature, *PID* controller tuning methods cover frequency response [9, 10], Internal Model Control, *IMC* [3, 11], Direct Synthesis [12], equating coefficient [13], stability analysis [14], optimization techniques [15], and integral error criterion [16]. On the other hand, *GA* has been applied to many different problems, such as: traveling salesman [17], graph partitioning problem, filters design, power electronics [18], machine learning [19], and dynamic control system [20].

The paper is organized as follows: Section 2 explained fitting on process and disturbance model, formulation of feedforward plus feedback algorithm, correlation tuning of *PID* controller, and principles of computational optimization analysis for feedforward plus feedback control scheme. Section 3 explained developed process and disturbance model, correlation tunings of *PID* controller and *GA,* which is validated through testing on Pumped-tank of LOOP-PRO software. The relative performance was presented in graphs, performance index and performance indicator. Finally, Section 4 presented conclusion of findings and analysis of applied tunings methods.

2 Formulation of Feedforward Plus Feedback Control Scheme and Stability Margin

The simplest way to determine the process dynamic behaviour of integrating process is by performing open loop bump tests; please refer to literature [2, 21]. In addition, using IMC-based formulas to determine correlation *PID* tunings are also well-explained in literature [22, 23].

Stability analysis is applied to obtain upper and lower limit of proportional gain, K_c and integral time constant, τ_I for *GA* optimization analysis. Refer to Fig. 1, the developed transfer function of feedforward plus feedback control scheme is shown in (1).

$$C = \frac{G_d + G_p G_{fc}}{1 + G_c G_p} D + \frac{G_c G_p}{1 + G_c G_p} R \quad (1)$$

The closed-loop stability is determined by its characteristic equation [24]. The closed-loop transfer function for load changes is explained in (2)

$$\frac{C(s)}{D(s)} = \frac{G_d + G_p G_{fc}}{1 + G_c G_p} \quad (2)$$

It's interesting to note that feedforward controller does not affect the stability of closed-loop control. The characteristic equation of disturbance is given as,

$$1 + G_c G_P = 0.$$

Applying Taylor approximation, $e^{-\theta_p s} \approx (1 - \theta_p S)$
Solve characteristic equation to obtain (3)

$$s^2\left(1 - K_c \theta_p K_p\right) + \left(K_c K_p - \frac{K_c}{\tau_I} \theta_p K_p\right) s + \frac{K_c}{\tau_i} K_p = 0 \quad (3)$$

From s^2 term, solve $1 - K_c \theta_p K_p > 0$ to obtain (4)

$$K_c < \frac{1}{\theta_p K_p} \; \textit{(Upper limit)} \quad (4)$$

From *s* term, solve $K_c K_p - \frac{K_c}{\tau_I} \theta_p K_p > 0$ to obtain (5)

$$\tau_I > \theta_p \; \textit{(Lower Limit)} \quad (5)$$

We do anticipate the perfect control, where the controlled variable maintains same at the setpoint eventhough arbitrary changes of the disturbance variable, D. Thus, the setpoint is constant (R(s) = 0), we want C(s) = 0 despite D ≠ 0, equation above is satisfied as

$$G_d + G_p G_{fc} = 0$$

Solving G_{fc} gives the ideal feedforward controller as shown in (6)

$$G_{fc} = -\frac{G_d}{G_p} \quad (6)$$

Dividing of G_d and G_p in Fig. 1, yields G_{fc} in (7)

$$G_{fc} = -\frac{K_d}{K_p} \, e^{-(\theta_d - \theta_p)s} \quad (7)$$

Integrating process does not have lead and lag element in feedforward controller's algorithm. Therefore, the feedforward controller only covers the ratio K_d and K_p. For feedforward controller to be realizable, the $\theta_d - \theta_p$ must be a non-negative. In the case of $\theta_d < \theta_p$, Erickson [24] suggested to choose θ_d to be similar value with θ_p so to make total deadtime equal to 0.

2.1 Genetic Algorithm for Measuring Integral Errors of Feedforward Plus Feedback Control Loop.

Fig. 2 shows the structure how is the error signals in both servo and regulatory control problems have been accumulated in *GA* optimization analysis. There are three types of minimum integral error signals were measured include Integral Absolute Error, *IAE*, Integral Square Error, *ISE* and Integral Time Absolute Error, *ITAE* index. The respective measurements are developed through Matlab as illustrated in Fig. 5.

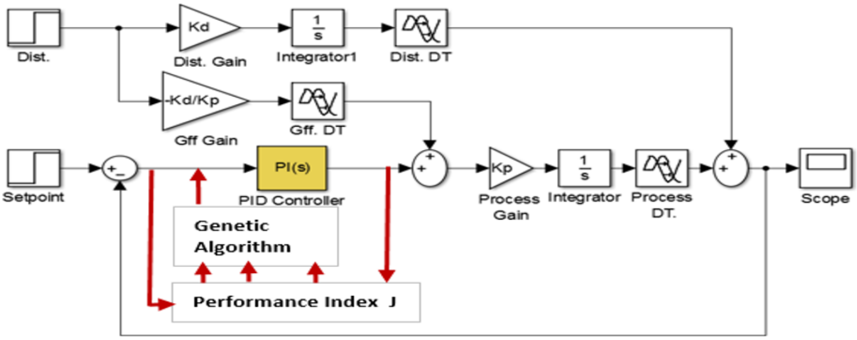


Fig. 2. Block diagram of computational optimization analysis using *GA*.

3 Analysis and Results

3.1 Process and Disturbance Model.

The *FOPDT-integrating* model for both process and disturbance are respectively determined through open loop bump test to the Pumped-tank of LOOP-PRO software. The actual $\theta_d \approx 0$. Therefore, we equalize θ_d with θ_p in forming *FOPDT-integrating* for process and disturbance bump test are shown in Table 1.

Table 1. Transfer function of *FOPDT-integrating* for process and disturbance bump test.

	Process	**Disturbance**
Transfer Function	$\frac{-0.0238e^{-0.9721s}}{s}$	$\frac{-0.0971e^{-0.9721s}}{s}$

3.2 Stability Margin.

From (4), substitute $K_c = 0.0239$ $\theta_p = 0.9721$ yields upper limit of K_c and re-state the range of K_c give $0 < K_c < 41.02$. Refer to (5), substitute $\theta_p = 0.9721$ yields lower limit of τ_i therefore we re-state the range of τ_i , by which $\tau_i > 1$.

3.3 PID Controller Tuning.

PI controller settings were applied to regulate the water level in the Pumped-tank. The correlation *PI* tuning values and feedforward gain, G_{fc} are tabulated in Table 2.

Table 2. PI controller and feedforward controller setting.

Tuning method	K_c	τ_I	Feedforward Gain, G_{fc}
PI Controller	-17.3	7.5	0
Feedforward plus feedback control (IMC)	-17.3	7.5	-4.063
Computational Optimization method (GA)	-33.79	6.62	-4.063

PI settings of feedforward plus feedback control scheme is similar to feedback-only control scheme. However, feedforward plus feedback control scheme has additional G_{fc}, which is the static ratio of K_d and K_p.

3.4 Improvement on Feedforward Plus Feedback Control Scheme in Regulatory Control Problem.

The transient and steady state responses of feedforward plus feedback control for Pumped-tank of LOOP-PRO software is illustrated in Fig. 3.

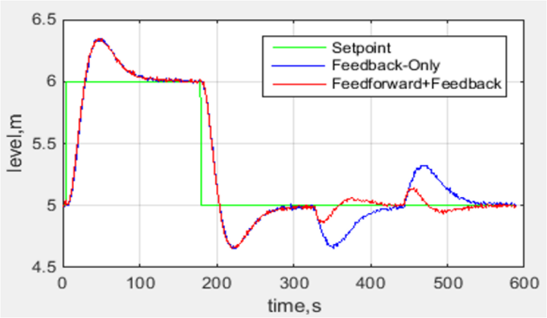


Fig. 3. Transient and steady state response of feedforward plus feedback and feedback-only control scheme.

From Fig. 3, feedforward plus feedback control scheme possess higher controllability to regulatory control problem as compared to feedback-only control scheme. It is noted that the added feedforward function has compensate the control action to process as the external disturbance to the process has changes.

3.5 Improvements for Both Servo and Regulatory Control Uses Genetic Algorithm (GA).

Fig. 4 illustrates relative performance of *GA* compared to other conventional tuning methods. Conventional feedforward plus feedback control scheme showed the improved performance in term of steady state response but have similar transient response as feedback-only control scheme. Besides, feedback-only control scheme in overall had performed poorly due to sluggish transient and steady state response.

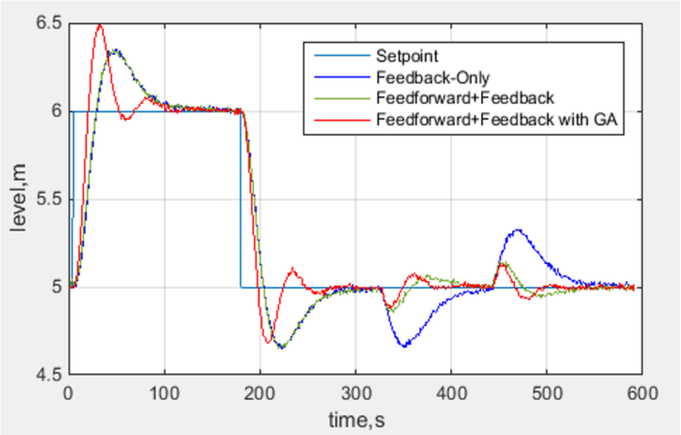


Fig. 4. Transient and steady state responses of feedback only, feedback plus feedforward and *GA* method.

Relative performance of Pumped-tank used *GA* optimization analysis had improved robustness of control actions thereby shorten the settling time as depicted in Fig. 4. Particularly for the regulatory control problem, *GA* produced less oscillated responses as compared to other tuning methods.

3.6 Performance Index

Overall performance of the system is evaluated by accumulating integral error signals of response through Simulink of Matlab software as illustrated in Fig. 5.

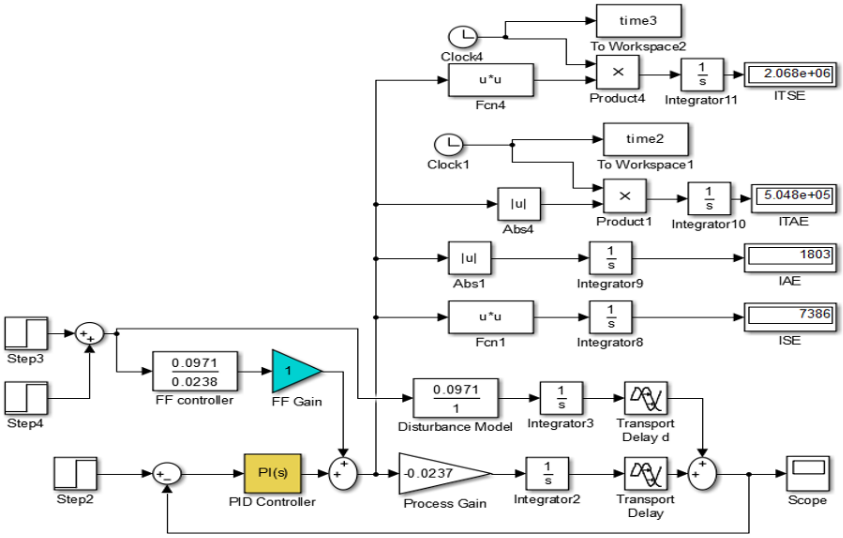


Fig. 5. Performance Index for feedforward plus feedback control scheme.

The respective servo and regulatory control problems were applied to the control loop then integral error values were recorded. All values were presented in indexes; whereby the smaller index value reflects the better performance. Performance indexes of feedback-only, feedforward plus feedback and *GA* were tabulated in Table 3.

Table 3. Performance index for servo and regulatory control problem

Tuning Method	Servo Control			Regulatory Control		
	IAE	ISE	ITAE	IAE	ISE	ITAE
Feedback-only control	14.86	3.445	635.5	1803	7868	5.05e+5
Feedforward plus feedback (IMC)	4.468	2.117	255.6	1803	7386	5.05e+5
Computational Optimization Analysis (GA)	3.169	1.858	169.6	1563	6404	3.02e+5

It is noted that computational optimization analysis for *PID* controller produced the lowest index values as compared to other tuning methods. *GA* tuning method produced the smallest error signals in term of IAE, ISE and ITAE. In contrast, feedback-only control scheme consequence the highest integral error signals.

3.7 Performance Indicator of Servo and Regulatory Control Problems

Performance indicator generally reflects how well the closed-loop system is performed for the applied *PID* settings and feedforward ratio. Relative performance indicator of feedback-only, feedforward plus feedback and *GA* is tabulated in Table 4.

Table 4. Performance indicator for servo and regulatory control problems.

Tuning Methodology	**Servo Control**			**Regulatory Control**	
	Rise time, s	Overshoot, %	Settling time, s	Overshoot, %	Settling time, s
Feedback-only control	24	35	91	23.3	69
Feedforward plus feedback (IMC)	24	35	88	10	20
Feedforward plus feedback (GA)	15	45	45	8.7	16

From Fig. 4, *GA* tuning method produced the shortest rise time and settling time as well as improved overshoots in regulatory control problem. It shows better performance compared to other methods therefore it was determined as optimal *PI* tunings for the tested function of LOOP-PRO simulation software.

4 Conclusion

The research is focused on the application of feedforward plus feedback control scheme in controlling an integrating process and studied on the improvements as compared to feedback-only control scheme. In addition, computational optimization analysis had been applied for finding the optimized *PI* tunings to the Pumped-tank function of LOOP-PRO software. Among computational optimization approaches, *GA* has the ability to consider for trade-off PI tunings that performs the best for both servo and regulatory control problems as shown in graphs, performance index and performance indicator. The best *PI* tunings for Pumped-tank function of LOOP-PRO software are K_c = -33.79 %/m, τ_i = 6.62 s and G_{fc}= -4.063.

References

1. D.Cooper, F.: Practical process control using LOOP_PRO software. Control Station, Inc. United State of America (2006).
2. R.Rice, F., D.Copper. S.: A rule design methodology for the control of non-self-regulating processes, https://pdfs.semanticscholar.org/231c/addd5ea6e56fee8aaa07086413c25008024e.pdf.
3. D.B.Santosh Kumar, F., R.P.Sree, S.: Tuning of IMC based PID controllers for integrating systems with time delay. ISA Transactions 63 (2016).
4. R.Kumar, F., S.K.Kingla, S., V.Chopra, T.: Comparison among some well-known control schemes with different tuning methods. Journal of Applied Research and Technology 13, 409-415(2015).
5. J.Nandong, F.: A Unified Design for Feedback –Feedforward Control System To Improve Regulatory Control Performance. International Journal of Control Automation and Systems, 1-8 (2015).
6. S.Padhee, F.: Controller design for temperature control of heat exchanger system: simulation studies. WSEAS Transactions on Systems and Control (9), 485-491(2014).

7. K.T. Erickson, F., J.L.Hedrick, S.: Plantwise Process Control. John Wiley and Sons Inc, United State of America (1999).
8. D.Cooper, F.: Practical process control using LOOP_PRO software. Control Station, Inc. United State of America (2006).
9. J.G.Ziegler, F., N.B.Nichols, S.: Optimum setting for automatic controllers. ASME Trans (64) 759-768 (1942).
10. G.H. Cohen, F., G.A.Cohen, S.: Theoretical consideration of retarded control. Trans ASME (75), 827-834 (1952).
11. D.E. Rivera, F., M.Morari, S., S.Skogestad, T.: Internal model control for PID controller design. Ind Eng Chem Process Des Dev 2(5); 252-265 (1986).
12. J. Lee, F., W.Cho, S., T.F.Edgar, T.: Simple analytical PID controller tuning rules revisited. Ind. Eng. Chem. Res (53), 5038-5047 (2014).
13. R.P.Sree, F., M.Chidambaram,S.: A simple and robust method of tuning controllers for integrator/dead time processes. J. Chem Eng. Jpn. 38(2), 113-119 (2005).
14. W.L.Luyben F.: Design of proportional-integral-derivative controllers for integrating/dead-time processes. Ind. Eng. Chem Res. 35(10), 3480-3483(1996).
15. A.Visioli, F., Q.C.Zhang, S.,: Control of integral process with dead time. Springer-Verlag, London Lmited (2011).
16. J.H.Holland, F.: Adaptation in Natural and Artificial Systems. Ann. Arbor, MI: Univ. Mich. Press (1975).
17. D.E.Goldberg, F., R.Lingle Jr, S.: Alleles, loci and traveling salesman problem. Proc. Int. Conf. Genetic Algorithms and Their Appl. 154-159 (1985).
18. B.Ozpineci, F., J.O.P.Pinto, S., L.M.Tolbert, T.: Pulse-width optimization in a pulse density modulated high frequency ac-ac converter using genetic algorithms. Proc. of IEEE System, Man and Cybernetics Conf. (3), 1924-1929 (2001).
19. J.H.Holland, F.: Genetic algorithms and classifier systems: foundations and future directions, genetic algorithms and their applications. Proc. of Sec. Int. Conf. on Genetic Algorithms (1987).
20. P.J.Van Rensburg, F., I.S.Shaw, S., J.D.Van Wyk, T.: Adaptive PID control using a genetic algorithm. Proc. KES'98, Second. Inter. Conf. Knowledge-Based Intel. Electro. Sys., (2). 133-138 (1998).
21. J.Smuts, F.: Level controller tuning, http://blog.opticontrols.com/archives/697.
22. P. Lee, F.: Tuning PID loops for level control, https://www.controleng.com/single-article/tuning-pid-loops-for-level-control/f2b4134403d7064939004ed946269ce7.html
23. B.Rice, F., D.Cooper, S.:A design and tuning recipe for integrating processes, https://controlguru.com/a-design-and-tuning-recipe-for-integrating-processes/
24. K.T.Erickson, F., J.L.Hedrick, S.: Plantwise Process Control", John Wiley and Sons Inc, United State of America (1999).

Performance Evaluation of Densely Deployed WLANs using Directional and Omni-Directional Antennas

Shuaib K. Memon[1], Kashif Nisar[2], Waseem Ahmad[3]

[1]Auckland Institute of Studies, New Zealand,
[2]Knowledge Technology Research Unit, Universiti Malaysia Sabah, Malaysia,
[3]Toi Ohomai Institute of Technology, New Zealand
shuaibm@ais.ac.nz, kashif@ums.edu.my,
waseem.ahmad@toiohomai.ac.nz

Abstract. It has been more than a decade since the Wireless Local Area Networks (WLAN) based on the IEEE 802.11 standard family has become commercialized. Inexpensive WLAN access points have found their ways in almost every household and enterprise and WLAN devices are embedded in the chips of laptops, tablets, mobile phones, printers, and many other household and commercial appliances. In recent years, there has been a tremendous growth in the deployment of IEEE 802.11-based WLANs under the brand name Wireless Fidelity (Wi-Fi). This growth is as a result of low-cost, international standard (e.g. 802.11a, b, g, n, and ac) flexibility and mobility offered by the technology. However, WLANs are deployed on a non-planning basis, not like public cellular phone networks. In WLAN, nodes commonly use omnidirectional antennas to communicate with access point. Omni-directional antennas may not be efficient due to interface caused by the transmission of packets in all the directions. Many researchers have evaluated the performance of Dense Wireless Local Area Network (a saturated network). There is need to evaluate the performance of densely deployed wireless networks (an area where high number of WLANs are deployed). This research paper has evaluated the performance of densely deployed WLANs using directional and omnidirectional antennas. Optimized Network Engineering Tool (OPNET) Modeler 17.1 has used to simulate directional and omnidirectional antennas.

Keywords: WLAN, WiFi, MAC, Directional Antenna, Antenna Pattern

1 Introduction

There has been a tremendous growth in the deployment of IEEE 802.11-based Wireless Local Area Networks (WLANs). The IEEE released the 802.11 standard for wireless LAN (WLAN) in 1997. The specification requires a data transfer rate from 1Mbps up to 2Mbps while retaining compatibility with existing LAN hardware and software infrastructure [1-2]. The standard defines protocols for Medium Access Control (MAC) layer and physical transmission in the unlicensed 2.4 GHz radio band. After successful implementation by commercial companies such as Lucent Technolo-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_36

gies, amendments were made for a better performance in the same year. The IEEE 802.11-based WLANs gained widespread popularity and become ubiquitous networks [3]. The MAC and PHY characteristics of 802.11 are specified in legacy 802.11-1997 [4]. And latter 802.11a [5], 802.11b [6] and 802.11g [7] PHY 802.11n, 802.11ac [8].

In typical office or shopping mall environment WLAN based infrastructures are used, where Access Point (AP) is fixed and nodes are mobile in the nature. By default, both AP and nodes use omnidirectional for transmission and reception. Omnidirectional antenna has high beam width, low cost and easy installation. Recently, there has been increasing interest in using directional antennas in 802.11 based WLANs. Directional antenna offers many benefits such as increased signal strength and transmission range, therefore it achieves high throughput. It also causes more deafness and hidden node problem. Therefore, there is a need to study the impact of directional antenna configuration in the context of network performance [10-11].

1.1 IEEE 802.11 WLAN Architecture

The fundamental building block of the 802.11 architecture is the Basic Service Set (BSS). A BSS contains one or more wireless Stations (STAs) and a central base station know as an Access Point (AP) in 802.11 parlances. Wireless LANs that deploy APs are often referred to as infrastructure wireless LANs, with the infrastructure being the APs along with the wired Ethernet infrastructure that interconnects the APs and a router as illustrated in Figure 1 and Figure 2.

IEEE 802.11 STAs can also group themselves together to form an ad-hoc network for the purpose of internetworked communications without the aid of an infrastructure network. An ad-hoc network is suitable for man-made or natural disasters such as U.S. 9/11, Boston Bombing, earthquakes where infrastructure networks were either not available or highly affected by the disaster.

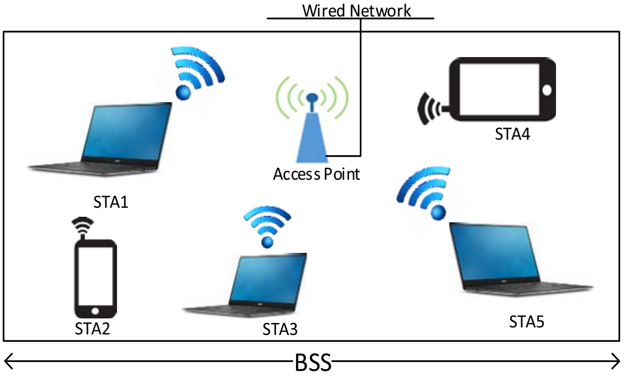


Fig. 1. A typical 802.11 wireless infrastructure network with five STAs.

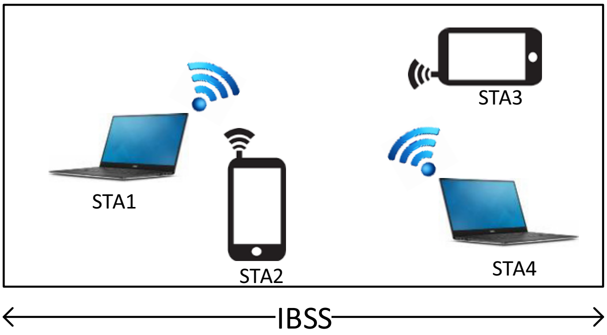


Fig. 2. A typical 802.11 wireless ad hoc network with four STAs.

2 Quality of Service

Quality of Service (QoS) is perceived and interpreted by different communities in different ways. The technical or network communities refer QoS as the measure of service quality provided by the network to the users. Internet Engineering Task Force (IETF) defines QoS as “a set of service requirements to be met by the network while transporting a flow” [11]. The main goal is to provide QoS while maximizing network resource utilization. The network user community refers QoS as the quality perceived by applications/users [12]. The International Telecommunication Union (ITU) defines QoS as “the ability of a network or network portion to provide the functions related to communications between users” [13].

3 Problem Statement

Quality of Services (QoS) is considered as the main issue in Wi-Fi connection management. Wi-Fi requires higher throughput, less delay and higher farness index over WLANs. The packets streaming can drop because of the competition among different kinds of traffic flow over the network. Therefore, the quality of internet users cannot be guaranteed. Wi-Fi traffic is also sensitive to delay and requires the packets to arrive on the time from sender to receiver side without any delay over the network. Therefore, current Wi-Fi networks are always slow.

4 Experimental Setup in OPNET Modeler

OPNET Modeler (network simulator tool) used to simulate the behavior and performance of any type of network. The main difference between OPNET Network Simulator to other simulators lies in its power and versatility. OPNET Technologies (including network simulators), build upon Riverbed's strong heritage of delivering industry-leading solutions to drive application performance. To assess and evaluate the performance of densely deployed wireless networks, we have created total 44 scenarios i.e. 11 scenarios each using directional antenna with 802.11a, and 802.11g, and 11 scenarios each using omni-direction antenna with 802.11a, and 802.11g. Figure 3 shows simple scenario of 802.11a (shows one AP and one node). However, Figure 4 shows dense wireless networks scenario of 802.11a (shows 100APs, one node is associated with each AP). Each access point has one node to measure the effect of mutual interference from other wireless networks.

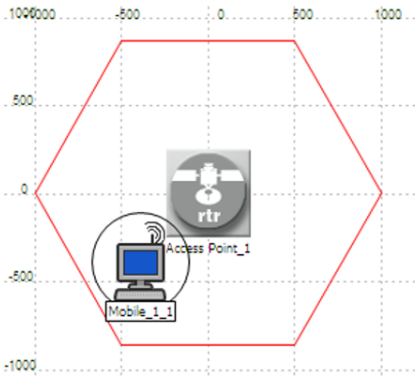


Fig. 3. Simple scenario

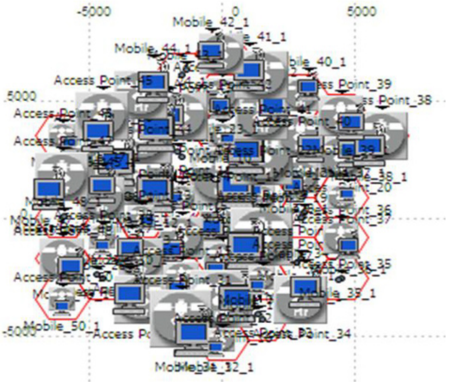


Fig. 4. Dense Scenario

Directional antenna module was implemented in OPNET Modeler. Next, WLAN node models were created for 802.11a and 802.11g networks by attaching an antenna module to the transmitter (Tx) and receiver (Rx) modules of the WLANs. Then, we set the antenna models up in the new antenna module to make it work successfully. Initially we verified Professor Umehira's (Ibaraki University, Japan) analytical models with our simulation results for an 802.11a network with omni-directional antenna.

We mainly focused on running extensive simulations, collecting and organizing simulation output data and graphical presentation of simulation results. For example, to run a simulation model of 802.11a with 50 APs, it took several hours to get the simulation output data. For network performance evaluation, we considered network throughput, packet delay, packet dropping and number of retransmission attempts. The results obtained show that high-density networks with a directional antenna perform better than traditional omnidirectional antenna.

We believe this a significant contribution to the field of high density networks, where network performance can be greatly enhanced by incorporating directional antenna.

5 Results

As per our results, Access points (APs) are networking devices that allow wireless Wi-Fi devices to connect to a wired network. They form wireless local area networks (WLANs). An AP acts as a central transmitter and receiver of wireless radio signals.

Mainstream wireless APs support Wi-Fi and are most commonly used in homes, to support public internet hot spots and in business networks to accommodate the proliferation of wireless mobile devices now in use.

We have performed extensive simulations to assess and evaluate the impacts of increasing Aps on the network performance. Figures 5 – 7 show the impacts of the number of access points on the throughput, delay and data dropped. Figure 5 highlights the effects of the number of AP on the average delay. We used 1 to 100 Aps (11 scenarios). It has been noticed that with the increase of access points, the average delay also increases. Using 10Aps, the average delay was less than 0.02, however, it increases to approximately 0.07 at 50 APs. Finally, when using 100APs, the average delay was 0.1.

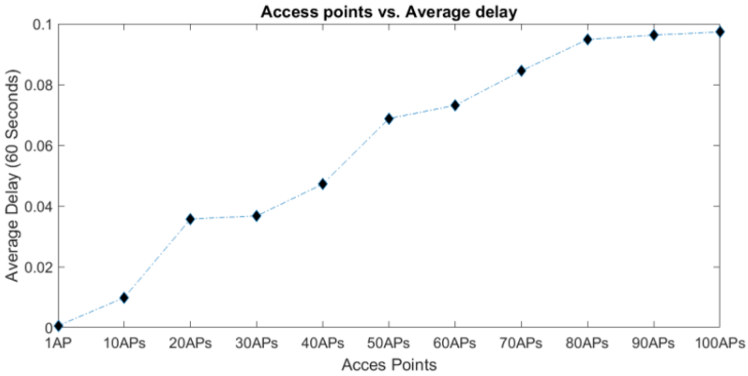


Fig. 5. Access points vs average delay

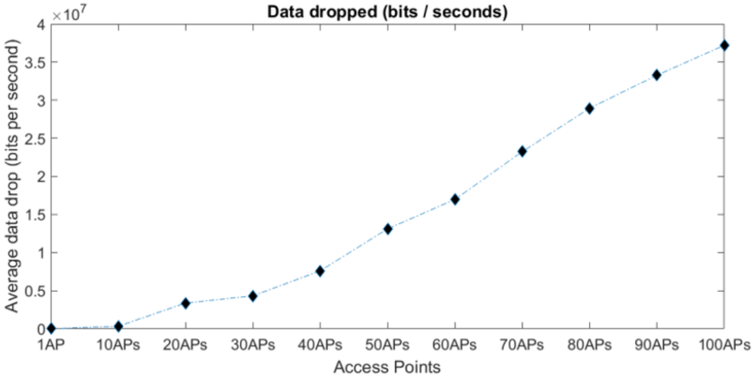


Fig. 6. Access points vs average data dropped.

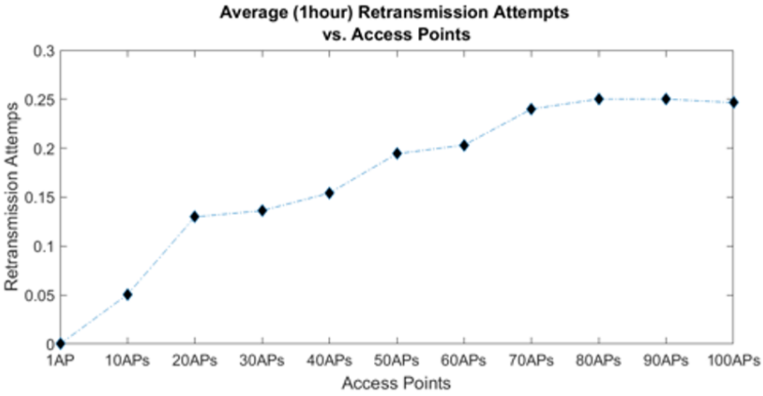


Fig. 7. Access points vs retransmission attempts.

Figure 6 demonstrates the relationship between access points and data dropped (bits/seconds). From 1AP to 10APs, there was almost no average drop in data. But from 10APs to 100Aps, the average data drop is significant. This suggests that APs and data drop has a positive relationship, with the increase in APs, average data drop increases. In Figure 7, we have plotted the access points vs average (1 hour) retransmission attempts. We have noticed that from 1AP to 100Aps, the number of retransmission attempt increases until 0.25. It can also be seen in Figure 7, that the number of retransmission attempt reaches a point of convergence after reaching approximately 70 APs. We have also assess and evaluate the impact of directional and omni-directional on the network. Figures 8 – 13 show the impact of direction and omni-directional antenna. Figure 8 shows throughput performance of directional vs omni-directional antenna. We used 100APs where, each AP was associated with one node.

Figures 8 and 9 illustrate the IEEE 802.11a and IEEE 802g throughput using direction and omni-directional antenna. Both simulations also show simulation time in seconds and throughput in bits/sec. Simulation time 0 to 60 seconds and throughput 0 to 100000000 bits/sec. In Figure 8, it can be noticed that directional antenna increased throughput from 10th second from the simulation time until the end of the simulation. As far as omni-directional antenna is concerned, it’s only increased at simulation time 3 to 6 seconds after that, it’s always less than directional antenna.

In Figure 9, it can be seen that both omni-directional and directional throughput increases rapidly with the increase of time in seconds, however, after 9 seconds the increase in throughput is steady.

Figures 10 and 11 illustrate IEEE 802.11a and IEEE 802g delay with direction vs omni-directional antennas. Both simulations show that after certain time period, directional antenna delay is decreasing, whereas, omin-directional antenna simulation has steady increased in delay. In figure 10, we have noticed that directional antenna increased delay from 0 second to until 18 seconds, after that delay started to decrease steadily. As for the ommi-directional antenna, delay was always increasing.

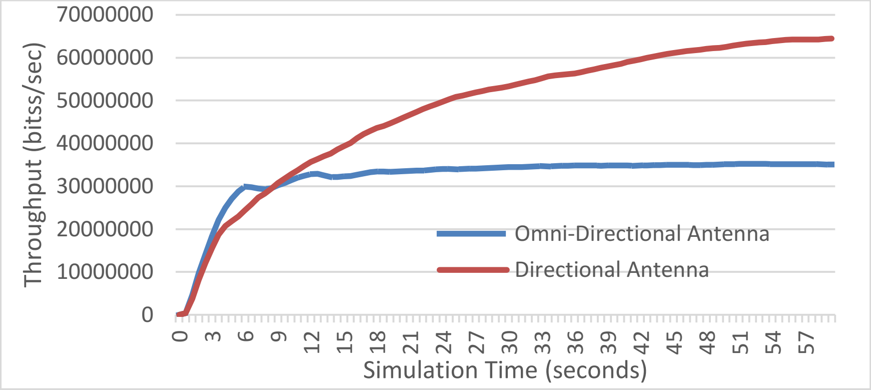


Fig. 8. 802.11a (throughput) direction vs omni-directional antenna.

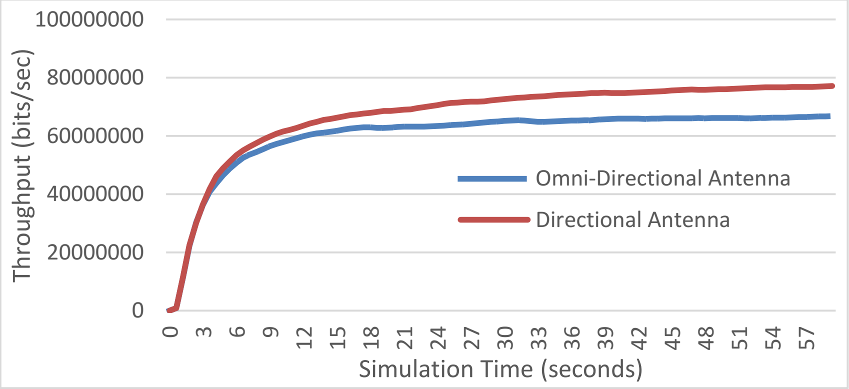


Fig. 9. 802.11g (throughput) direction vs omni-directional antenna.

Figures 12 and 13 illustrate the IEEE 802.11a and IEEE 802g data dropped with direction and omni-directional antennas. Both simulations (in Figure 12 and 13) are showing time in seconds and data dropped in bits/sec. In Figure 12, it can be seen that that data dropped started to decrease in directional antenna from 8th second until the end of simulation. However, for ommi-directional antenna, the data dropped rate was always increasing and higher than directional antenna simulations. In Figure 13, we can observe trends similar to Figure 12, where directional antenna simulations results are superior to omni-directional antenna simulations.

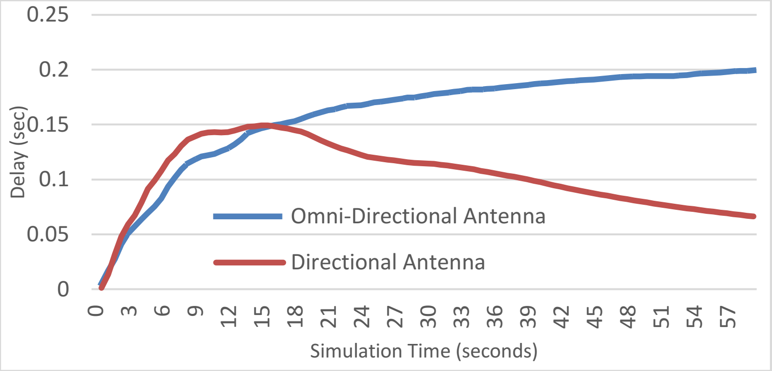


Fig. 10. 802.11a (delay) direction vs omni-directional antenna.

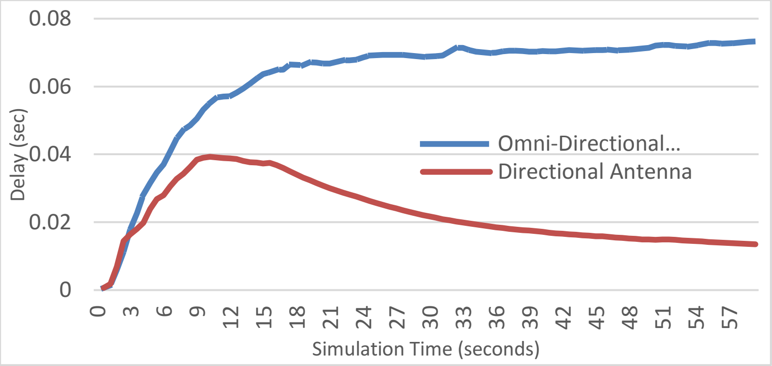


Fig. 11. 802.11g (delay) direction vs omni-directional antenna.

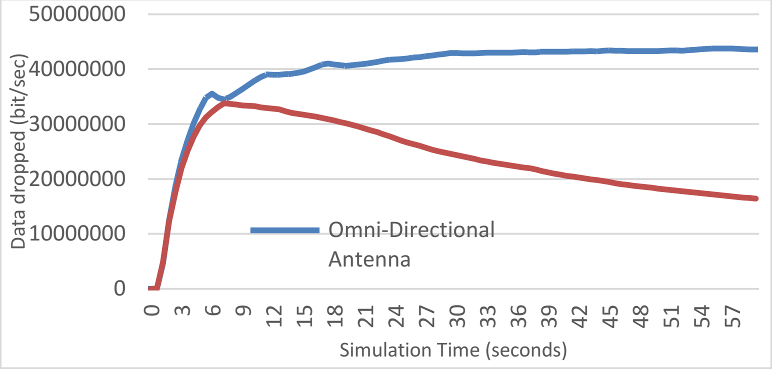


Fig. 12. 802.11a (data dropped) direction vs omni-directional antenna.

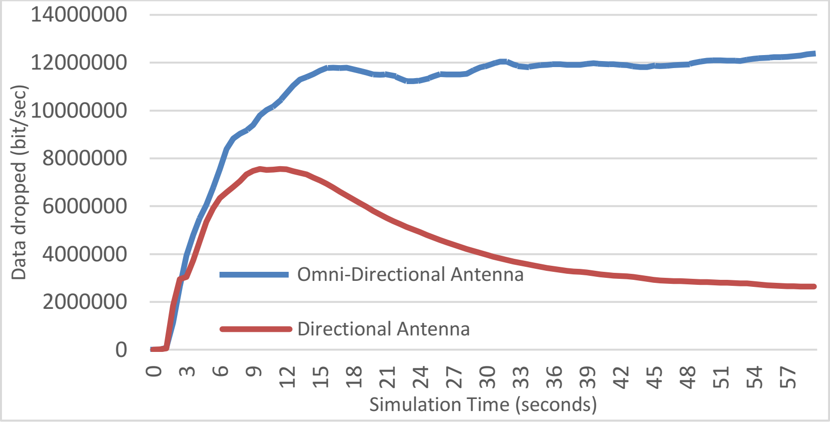


Fig. 13. 802.11g (data dropped) direction vs omni-directional antenna.

Table 1. Parameters used

Parameters	Status
• Start Time (Seconds)	Constant (1)
• Off State Time (Seconds)	Constant (5)
• Off State Time (Seconds)	Exponential (1)
• Packet size	250 exponential
• Network	802.11a and 802.11g
• Traffic	data
• Performance Matrices	data dropped, delay and retransmission attempts

6 Conclusion & Future Work

Performance of network is highly affected in dense environment, where increase in the number of APs decreases the network performance. In this research paper, a thorough comparison of omni-directional and directional antennas was performed. The experimental results concluded that directional antenna I was expecting similar delay, data dropped or retransmission attempts between 10APs to 20APs, 20APs to 30APs, 30APs to 40APs, 40APs to 50APs and so on, but it is not. We have used packet size 250 Bytes, this is still ideal packet size in highly dense scenario as delay, data dropped or retransmission are concern, there will be more data dropped, packet delay, and retransmission attempts if we increase the packet size in dense networks. Packet Size 1024 is suitable in scenario has 20AP as compare to Smaller Packet Size (Packet Size 250). If APs are increased from 20 then small packet size is better. In future, we will implement our results on Testbed over WLANs.

References

1. L. Romdhani, N. Qiang, and T. Turletti, "Adaptive EDCF: enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks," in *Proceedings IEEE Wireless Communications and Networking Conference*, 2003, pp. 1373-1378 vol.2.
2. W.-Y. Lin and J.-S. Wu, "Modified EDCF to improve the performance of IEEE 802.11e WLAN," *Computer Communications,* vol. 30, pp. 841-848, 26 February 2007.
3. O. Shagdar, K. Sakai, H. Yomo, A. Hasegawa, T. Shibata, R. Miura*, et al.*, "Throughput maximization and network-wide service differentiation for IEEE802.11e WLAN," in *International Conference on Communications and Information Technology (ICCIT), 2011*, 2011, pp. 43-46.
4. K. Kosek-Szott, M. Natkaniec, and A. R. Pach, "A simple but accurate throughput model for IEEE 802.11 EDCA in saturation and non-saturation conditions," *Computer Networks,* vol. 55, pp. 622–635, February 2011.
5. T. Sanada, X. Tian, T. Okuda, and T. Ideguchi, "Estimating the Number of Nodes in WLANs to Improve Throughput and QoS," *IEICE Transactions on Information and Systems,* vol. 99, pp. 10-20, 2016.
6. I. Syed, S.-h. Shin, B.-h. Roh, and M. Adnan, "Performance Improvement of QoS-Enabled WLANs Using Adaptive Contention Window Backoff Algorithm," *Journal of IEEE Systems* vol. PP, 2017.
7. S. Choi, J. Prado, N. Shankar, and S. Mangold, "IEEE 802.11 e contention-based channel access (EDCF) performance evaluation," in *IEEE International Conference on Communications*, 2003, pp. 1151-1156.
8. Y. Xiao, L. Haizhon, and C. Sunghyun, "Protection and guarantee for voice and video traffic in IEEE 802.11e wireless LANs," in *The 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2004, pp. 2152-2162 vol.3.
9. Y. C. Lai, Y. H. Yeh, and C. L. Wang, "Dynamic Backoff Time Adjustment with Considering Channel Condition for IEEE 802.11e EDCA " *Information Networking,* vol. 5200, pp. 445-454, 2008.
10. W. Jian-xin, S. MAKFILE, and J. Li, "A random adaptive method to adjust MAC parameters in IEEE802.11e WLAN," *Wireless Networks,* vol. 16, p. 629−634, 2009.
11. E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "A Framework for QoS-based Routing in the Internet," Network Working Group: Request for Comments: RFC - 2386Aug. 1998.
12. "ITU-T Telecommunication Standardization Sector of ITU," in Telephone Networks and ISDN: Quality of Service, Network Management and Traffic Engineering, ed: ITU-T E.800, 1994.
13. ITU-T, "Series E: Overall network operation, telephone service, service operation and human factors," Sep. 2008 2008.

Analyzing National Film Based on Social Media Tweets Input Using Topic Modelling and Data Mining Approach

Christine Diane Ramos, Merlin Teodosia Suarez and Edward Tighe

De La Salle University, 2401 Taft Ave, Malate, Manila, 1004 Metro Manila
christine.diane.ramos@dlsu.edu.ph,
merlin.suarez@delasalle.ph, edward.tighe@dlsu.edu.ph

Abstract. This paper presents a methodology to measure and analyze mass opinion towards underdeveloped forms of art such as independent films using data mining and topic modelling approach, rather than limited sampling through traditional movie revenue and surveys. Independent films are cultural mediums that foster awareness and social transformation through the advocacies and social realities they present. This methodology helps in addressing challenges of film stakeholders and cultural policy-making bodies in assessing cultural significance of independent films. Twitter has allowed innovative methods in data mining to understand trends and patterns that provide valuable support in decision making to domain experts. By determining the status of Philippine Cinema using social media data analytics as the primary source of the collective response, film stakeholders will be provided with a better understanding how the audience currently interprets their films with results as quantitative evidence. We use the tweets from the Pista ng Pelikulang Pilipino given the festival objective of showcasing films that enhances *"quality of life, examine the human and social condition, and contribute to the nobility and dignity of the human spirit"*.

Keywords: Twitter, Data Mining, Topic Modelling, Film

1 Introduction

1.1 Cultural Significance of Philippine Cinema

One of the innovative projects to address human development challenges under the United Nations Development Programme is the promotion of documentaries and independent films as an indirect persuasion to address social realities and issues [1]. Documentaries or independent films are cinematic publications that bring viewers to new worlds and experiences in the presentation of information about real people challenges, places, and events [2]. Most aim to reach out to the community to raise awareness and educate them on realities that explore issues on social, political, religious and psychological landscapes. The effect of how the audience can make emotional realizations by understanding an issue in a whole new perspective and expressing it in social media as a public opinion becomes a strong tool behind social trans-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_37

formation and mass enlightenment[3]. The motive for conducting this study stems from the fact that there has been no research in the movie domain that have analyzed movie goer reactions to understand and explain cultural behavior patterns and trends. Several studies surrounding the movie domain have been centered on predictability of box office sales based from relationships drawn out from its surrounding technical variables such as number of famous actors in the film, famous director, genre, etc. This gained focus more on the technical aspects of film rather than the analysis of content and meaning it imparted to the audience. The challenge is to dig deeper into social media streams to capture and assess the content generated by users, and to capture elicitation of such content in achieving a better understanding of the movie-goers perception of independent film [4]. To independent directors and stakeholders, audience perception in social media is necessary especially that Twitter streams are forms of "progressive accumulation of shared knowledge". Thus, the conversational patterns generated by the different movie goers may them aid in their film production strategies.

Social media provides an opportunity of abundant unsolicited data opinion interaction patterns from a wide social network, with tweet data sets ranging from thousands to million tweets across several social media mining studies [5]. A large amount of user interaction formation allows analysis of dynamic sentiments of users on different topics to investigate realistic opinions. Opinion in social media presents an innovative opinion model that explores how individual behavior affects a collective phenomenon [6]. This study focuses on the analysis of the contents generated by select Twitter users, specifically in obtaining insight from their reactions towards the independent films showcased in the Pista ng Pelikulang Pilipino 2017 festival.Social media, more than a platform for promotional strategies, may be also used as a more abundant authentic source to understand and explain the cultural behavior on how the Filipino audience perceive independent films vs. limited sampling and low responsiveness in surveys. To appeal to the segment of the masses, there is a need to advocate better concepts for films, especially that it constitutes the backbone of social realities [7]. By determining the status of Philippine Cinema using social media analytics as the primary source of the collective response, film stakeholders will be provided with a better understanding how the audience currently interprets their films using data mining and analytics as quantitative evidence.

1.2 Research Questions

In the Philippines, though several presidential decrees or policies have been created as an expression of gratitude towards national artists, there is still absence of national arts appreciation as this is compounded by the lack of interest of the community in the potential role and relevance of non-profit artists such as independent film makers in the contemporary arts fields[8]. Currently, the National Commission for Culture and the Arts (NCCA) included in this year's thrust the call for innovations and new technologies in the art. NCCA is the highest policy making arts council in the Philippines. It is responsible of coordinating grants and formulating policies for the preservation, development and promotion of Philippine arts and culture. The NCCA envisions culture as the "core and foundation of education, governance and sustainable development." [10] In support of NCCA's main goal, this study proposes social media opinion mining as an innovative means to identify emerging societal trends based on movie goer views, moods, attitudes, and expectations towards interpreting meaning with independent films. A major application of social media opinion research is the area of policy making to better understand real-world observations of communication patterns and anticipate likely impacts of certain issues or topics [1]. The democratization of web publishing has led to the significant increase in number of opinions expressed over the internet which allows citizens to be more actively engaged and empowered, sometimes demanding action and recommendations from perspectives presented in the online community [11]. Thus, this study aims to answer the following research questions; 1) What are the resulting themes that drive social media reactions after watching PPP2017 independent films? 2) What insight can be drawn to explain the societal interpretation of how independent films are perceived?

2 Related Works

Contrary to other social media platforms, Twitter is a form of participatory social media platform that allows users to engage in the creation and distribution of its content by sharing, receiving, and posting of tweets up to 140 characters in length but with hyperlinks, users can share more substantial information [11]. In the movie domain, data mining and machine learning using twitter data are grounded on introducing predictive algorithm models for film predictability, none on social behavior analysis. For instance, Lash and Zhao introduces a Movie Investor Assurance System framework aims to capture accurate forecasts on movie profitability referencing from previous historical data surrounding who are involved in the movie, what the movie was about, when the movie was release. This was evaluated with different criteria of profitability [12]. The study of Apala, et al. used Twitter, You Tube and IMdb movie databases to create a model of both uniform and non-uniform weighted approach to assess if a movie was a flop, a hit or neutral based from variables such as genre, director, and actor popularity. Results concluded that actor popularity was a significant variable along with movies that have a successful prequel [13]. The same study was also applied by Demir, et al. wherein the IMBD movie rating platform was explored to predict movie profitability based from quantitative indicators such as the number of views, likes, comments and favorites. A mathematical model was formulated to see correlation between likes over dislikes and linear regression for cross validation [14]. Another study of Chong Oh, et. al. further examined the perspective of the consumer engagement behavior among social media platforms such as Facebook, You Tube, and Twitter. This study proposed metrics to associate the consumer engagement behavior with correlations between factors such as count of Facebook likes, Follows in Twitter, and Views in YouTube as predictability on movie gross revenue [15]. Another perspective on social media analysis explores the role of the source of information and the number of its followers. The primary source with high follower reach is the effective indicator of movie sales on movies [16]. Other research such as that of

Mukhopadhyay, et. al. would explore sentiment analytics to assess the writing style of the community in responding to a film- if words used are mostly negatively or positively phrased [18]. To our knowledge, no research to date have used machine learning for knowledge discovery using public opinion for film analysis. Though profitability is necessary for more sound investment decisions, analyzing reactions are also important for cultural policy stakeholders to assess significance of film in terms of viewer affect and meaning. Gao, et al.'s is a similar research that explored analyzing film through the analysis of reviews posted in IMDB but the intent of this paper was to track differences of reviews based from social voting mechanism [17]. Gao's topic classification was used as the baseline reference coding scheme for this study. More than a microblogging platform on data mining for revenue prediction, the role of social media data paves emergence of a collective identity to coordinate and take action. Twitter data mining is significant especially to researchers, policy makers and government given how fast information in the internet is produced and diffused to the public at large[19]. Leavey highlights the capability of social media such as twitter to improve the quality and timeliness of the evidence base that informs public policy. The ability of human connections and interactions to provide insight, enabling various government units assist to advocate relevant and timely policies in their respective domains. Key findings from case studies allowed corroboration and verification of social media data when matched to other data sources [20]. In contribution to the study, this research presents the opportunities on the influence of twitter data mining towards policy making in culture and the arts as an attempt to address the current issue on the inability of producing quantitative evidence in conjunction to the cultural policies proposed [9]. In the study of Bautista and Lin, content analysis was done during and after the National Day of Mourning. The coding scheme was guided by Cutrona and Suhr's social support behavior codes in classifying social support responses in five categories- information, assistance, active participation, esteem, and advice. Results yielded four tweet categories- informational support (posting and resharing of information about the event), emotional support (paying tribute, sympathy, prayers, and expressions of grief), and non -social support (spams, anger, humor). Informational support has the highest number of tweets compared to the other categories.. Though SNS may be a platform for facilitation of social support, it can also be a platform for malicious messages and hyperlinks taking advantage of the situation [21]. Another social research was that of Soriano, et al in their study that explored the types of citizen engagement during the Yolanda typhoon. Contrary to the traditional one-way interaction presented in media, social networking sites such as Facebook and Twitter have become forefront tools in participatory media which allows consumers not just to receive information, but also create and distribute them. Using content analysis and topic modelling, disaster tweets were identified such as Solidaristic or Answerability/Responsibility [22].

3 Methodology and Findings

The proposed methodology was inspired by the study of Lippizi, et. al. on their data mining study in proposing models on conversation combinations, online traffic, sentiment and social analytics [2], combining with the content analysis approach in the study of Soriano, et, al.. on social media and civic engagement from analyzing tweets during calamities using data mining and topic modelling [12]. The study uses a new approach to policymaking particularly for government department on the arts. The domain is under-explored, and leverages on voluminous and available data to automatically craft themes and knowledge to support policy makers. It uses qualitative methodology as an empirical assessment based on content analysis and follows the epistemology of constructivist/ interpretivism approach to develop a truth based on the social interactions.

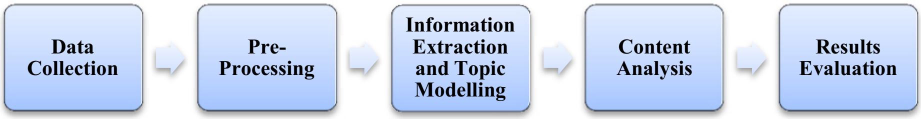


Fig. 1. Research Methodology

3.1 Data Collection

The collection of data was performed using a Python script interfacing with Twitter's Streaming API. The collection process adheres to the Twitter User Agreement which allows free Twitter Streaming comprising of 1% of total tweets belonging to public profile accounts. To pinpoint and monitor the tweets related to the Pista ng Pelikulang Pilipino 2017, specific keywords were selecting based on the official hashtags used by each movie. Official hashtags of the PPP 2017 were also included. A total of 146,214 tweets was collected from August 4, 2017 to September 8, 2017. Further information, such as Time Created, Text, Language, and Hashtags, was then extracted from the metadata of each of tweet. The data attributes extracted are the following: id, created_at_PHT, text, lang, is_a_retweet, is_a_reply, is_a_quoted_tweet, hashtags, urls, and user_mentions. It is interesting to note that upon manual inspection of the tweets, a sizeable cluster was identified to have been using one of the included keywords (#AWOL), but had no direct relation to the actual movie. Removal of unrelated tweets was performed by searching all tweets with the keyword "AWOL" and any combination of frequently occurring keywords A total of 20,251 unrelated tweets were identified and removed bring the corpus size to 125,963.

3.2 Pre-Processing

Processing the entire corpus would give a view of the entire film festival, but grouping tweets according to movies would allow for a more focused analysis. Given a movie group, natural language processing could output a representation specific to the movie and possibly reduce the amount of noise attributed by mixing discussions specific to other movie groups. Therefore, tweets were clustered according to each movie's keywords. If a keyword (without the hash symbol) was found in a tweet's text, then it would be grouped with all other tweets that contained the keyword. A tweet could belong to multiple movie groups for as long as the keyword appeared. As there were twelve movies, a total of twelve groups were created. The movie groups and their respective number of tweets can be found in Table 1.

Table 1. PPP 2017 Text Corpus

Movie Titles of PPP 2017 Films	Genre	Count
100TulaParaKayStella	Drama/Romance	37,890
PatayNaSiHesus	Drama/Comedy	8,567
AWOL	Action	9,186
Barboys	Drama	7,340
Triptiko	Romance	7,336
Birdshot	Thriller	6,743
Salvage	Horror	1,558
PauwiNa	Drama	2,143
Paglipay	Romance	568
StarNaSiVanDammeStallone	Comedy	531
Hamog	Thriller	418
AngManananggalSaUnit23B	Romance	292
Total		**125,962**

All tweets were then processed in order to turn simple characters into useful information. The text from each Tweet was extracted and tokenized using Tweetokenize, an external Python library that can not only identify and separate words, but also recognize social media entities such as user mentions (@usernames), hashtags (#hashtag), and URLs. The following rules were then set during tokenization:

- Words were lowercased, except when composed of all capital letters
- Normalization was set to 3 characters (e.g. *hmmmmm* is reduced to *hmmm*)
- Identified user mentions, hashtags, and URLs were left in their raw text form – no manipulation was applied
- Stop words in both English and Filipino were removed

3.3 Information Extraction and Topic Modelling

Information Extraction is the creation and manipulation of relations extracted in performing text structuring. This research approaches extracting information from a vast amount of unstructured data by getting the top occurring words and applying topic modeling on each of the movie group. The first approach looks at sifting through a large amount of data and coming up with the most common words used across tweets. To extract the top occurring words for a given movie group, a bag-of-words unigram model was constructed based on the group's tokenized tweets. Top occurring words are then defined as tokens with the highest term count within the movie group. The second approach looks to understand how words occur together. Topic modeling is applied to automatically discover clusters of words that are associated with each other because of their frequency of occurrence across tweets. This research utilizes Non-Negative Matrix Factorization (NNMF) to extract topics. Term Frequency Inverse Document Frequency (TFIDF), an extraction method that applies weighting scheme that scales up rare words and scales down too common words, is first applied to the bag of words model of a given movie group. This produces a term-document matrix A of size $m \times n$, where m is the number of terms and n is the number of documents or tweets. Given k topics, NNMF is the process of decomposing A into a feature-topic matrix W and a topic-document matrix H, where

$$A_{m\times n} \approx W_{m\times k}H_{k\times n}\#(1)$$

NNMF was implemented using Scikit-learn [22] where $\boldsymbol{k}$ was set to 30. The top words for each topic, as well as their respective rank within a topic, are returned to better understand the association between the words in the cluster and for comparison with coding themes.

3.4 Content Analysis and Results

Topic classification were mapped based from expert judgment following content analysis approach similar to the study of Soriano, et. al. [22]. The word clusters produced by the topic model algorithm were mapped to the coding themes developed by Gao, et. al. in classifying movie review reactions [17]. This was validated using Cohen's Kappa Coefficient where all are found to have a significant and strong Kappa's score average (κ= .846, p=0.00; κ= .752, p=0.00; κ= .794, p=0.00). Top frequent terms were also considered in validating analysis on the topic classification. Table 2 shows a summary of the topic models, frequently occurring terms, and topic classification of each movie.

Table 2. Topic Models and Classification Per Movie

Movie/ Synopsis	Word Clusters (rank)	Frequently occurring terms (count)	Topic Classification
100TulaParaKayStella (100 poems for Stella) *Fidel and his crushie Stella go through college together and we watch as he struggles to confess his feelings for her.*	😭 (10.24), feels (0.17), loves (0.14), ganda (0.13), grabe (0.12), pinaiyak (0.11), sobrang (0.1), 😚 (0.09), manood (0.09), 💕 (0.09), ❤□❤□❤□ (0.09), iyak (0.09)	RT 13540), 😭 (6777), @padillabela (6497), … (4747), mo (4549), fidel (3861), @qaloyJCsantos (3672), stella (3343), yung (3239), 💔 (3175)	Feelings towards the Film
PatayNaSiHesus (Jesus is dead) Jaclyn Jose plays mother Iyay, who takes her *kids on a road trip, to attend the funeral of her estranged husband.*	kaayo (0.39), lingaw (0.39), #pistangpelikulangpilipino (0.11), 😂 (0.44), bet (0.32), ating (0.32), panonoorin (0.32), mapanuod (0.32), talaga (0.32), movie (0.29), worth-it (0.03), tara (0.03), pnpp (0.03), makapanuod (0.03)	RT (194), #PPP2017 (111), #PistaNgPelikulangPilipino (105), tickets (97), tula (97), napanuod (97), 100 (97), buy (97), #PlayItRight (97), @GMoviesApp (97)	Film Endorsement*

AWOL *Abel Ibarra goes after the man who tried to kill him and his family.*	@starcinema (3.14), cinemas (2.77), nationwide (2.05), watch (1.67), rt (1.31), #pppextendedsacinelokal (0.88), family (0.81), protect (0.67), wag (0.65), papalampasin (0.61), nood (0.51), select (0.35), kapamilya (0.3)	RT (7594), … (2036), #PistaNgPelikulangPilipino (1618), gerald (1529), @CeciPhantomhive (1407), @StarCinema (1258), anderson (1180), @splatout76 (898), cinemas (752), @AarDooo (631)	Film Endorsement (mention of famous actor)*
Barboys *Four law students and the sacrifices they make as they inch closer to their dream of passing the bar and becoming lawyers.*	encouraged (1.57), evidence (1.57), student (1.55), support (1.5), law (1.47), enlightened (0.16), youth (0.16), tomorrow (0.14), degree (0.08), temporary (0.08), survive (0.06), pain (0.06), represents (0.05)	RT (4297), @nacinorocco (1482), love (1166), #PistaNgPelikulangPilipino (1110), … (1054), watch (813), @Enzo_Pineda (744), movie (634), 😂 (513), https (494)	Relate to Personal Experience/ Lessons Learned on the Film
Triptiko *A playboy who runs out of luck, a model whose skin-deep curse reaches a boiling point, and a musician chasing after light-footed love*	natin (3.29), susuportahan (1.67), pasaway (1.67), sobrang (1.57), parin (1.43), dun (1.27), kaibigan (1.14), mo (2.87), yung (2.2), kaibigan (2.05), nandyan (0.95), oras (0.67), mahal (0.56), palagi (0.38)	RT (5310), @RicciRivero06 (2982), mo (1075), yung (908), watch (825), trip (737), @PatamaDiary (716), kaibigan (684), (678), … (665)	Relate to Personal Experience/ Lessons Learned on the Film
Birdshot *After unintentionally shooting a Philippine Eagle, Maya is forced to flee in a forest where threats lurk in the dark.*	amazing (1.58), cast (1.52), story (1.52), cinematography (1.5), birdshot (1.45), screens (1.17), #birdlife (1.09), #naturephotography (0.85)	RT (5017), … (4246), @BirdshotPH (3636), #PPP2017 (3176), #TheDebateWithin (2889), create (2868), manood (2865), di (2829), https (2777), eh (2767)	Technical Details on the Film
Salvage *News crew reporting on alleged mythical creatures in Mindanao are chased by an armed group into the forest.*	vintage (0.67), @etsy (0.62), metal (0.56) #luxury (0.8), #exotic (0.79), #germancars (0.76) horror (0.72), filipino (0.62), film (0.59), disturbing (0.26)	… (646), RT (534), #forsale (172), #PPP2017 (149), https://t.co/cT79ehasTj (138), #vintage (137), #PistaNgPelikulangPilipino (99), #exotic (93), #luxury (90), salvage (89)	Other usage of hashtag* (majority) Technical Details on the Film
PauwiNa *A sickly man, his nagging wife, a blind and pregnant woman, "Jesus Christ", and a dog ride a pedicab out of Manila to find greener pastures in the province.*	masuportahan (1.81), neysyen (1.81), panuorin (1.79), sana (1.77), nyo (1.73), salamat (1.73), recommend (0.99), watched (0.91), magpromote (0.79), nyo (0.79), ganito (0.79), mapapanuod (0.79)	#PauwiNa (1956), RT (1255), ... (933), … (902), #PistaNgPelikulangPilipino (631), @IamJNapoles (587), po (445), nyo (425), love (411), salamat (402)	Film Endorsement

Paglipay *An Aeta crosses the river to go to town and find a wife. On the way he meets a woman who grew up in the city and the encounter creates a lasting impact*	portrays (0.52), wonderfully (0.52), gorgeous (0.52), thought-provoking (0.52), ♡ (0.52), aeta (0.49), paglipay (0.47) beautiful (0.46), culture (0.45)	RT (333), … (187), #PauwiNa (150), #PistaNgPelikulangPilipino (105), (98), #PPP2017 (92), 👍 (61), recommend (54), #AMU23B (49), watch (46)	Social Reality/Promoting Culture*
StarNaSiVanDammeStallone *A child with Down Syndrome dreams of being a start*	❤ (0.6), ☺ (0.53), solid (0.49), must-watch (0.48), 🎥 (0.48), 👋 (0.47), films (0.44) guys (0.43), tagos (0.39) puso (0.39)	RT (309), #PistaNgPelikulangPilipino (218), … (189), 4 (95), #PPP2017 (91), #PauwiNa (67), 2 (60), watch (60), 5 (60), 3 (54)	Feeling towards the film / Film Endorsement*
Hamog Four kids try to survive the mean streets of Manila	alcantara (0.8), starring (0.79), kyline (0.71), @starmagicphils (0.69), jaranilla (0.69), zaijan (0.68)	RT (242), … (200), #PistaNgPelikulangPilipino (99), @IamLavinia (70), film (61), watch (56), @Theresitaaaa (56), #PauwiNa (50), #PPP2017 (44), SM (40)	Film Endorsement (via actor/producers)
AngManananggalSaUnit23B *A character who has to keep emotions at bay or risk transforming into a winged creature.*	kang (0.84), wala (0.84), pag-ibig (0.41), pinipili (0.41), napakapure (0.41) hinuhusgahan (0.41), naman (0.32)	RT (178), #PistaNgPelikulangPilipino (125), … (92), @iamryzacenon (55), #AMU23B (54), #PPP2017 (45), mo (27), #pistangpelikulangpilipino (26), 3 (24), 2 (23)	Lessons of the Film/ Relating to Personal Experience

4 Conclusions and Future Development

6 out of the 11 movies from PPP2017 was able to invoke meaning and affect from the social realties they presented especially "Paglipay", where it was the only movie reactions centered on the Aeta culture appreciation. Unfortunately, this movie did not gain any award. With this data, stakeholders may consider a new form of metric criteria upon assessment films for award selection. This methodology considers factual evidence from mass opinion more than expert judgment or opinion editorials. Emerging themes* such as tweets on Film Endorsement, Promoting Culture, and Other Usage of Hashtag were also discovered. This highlights importance of hashtag usage to pull correct data pertaining to film. For instance, "Salvage" is a common term for car vehicles for re-selling. Thus, results yielded topic models mostly on car discussions vs. the movie itself. In this research, it was difficult to differentiate lessons learned from personal experience encountered and this needs further content analysis which will be good for future works. Future research may also want to consider the user as the unit of analysis.

Acknowledgements

The authors would like to thank NCCA for providing funding assistance in fulfillment of this research and the reviewers who have provided insightful comments on the first version.

References

1. United Nations Development Programme, "About Human Development," *2015.* .
2. S. Bernard, *Documentary Story Telling*. Elsevier Inc., 2011.
3. B. Mendoza, "Philippine Indie Films Make Headway," *The Philippine Star*, Mar-2009.
4. C. Lipizzi, L. Iandoli, and J. E. R. Marquez, "Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies," *Technol. Forecast. Soc. Change*, vol. 109, pp. 35–49, 2016.
5. F. Xiong and Y. Liu, "Opinion Formation on Social Media: An Empirical Approach," *An Interdiscip. J. Non-Linear Sci.*, 2014.
6. M. Kaschesky, P. Sobkowicz, and G. Bouchard, "Opinion Mining in Social Media: Modeling, Simulating and Visualizing Political Opinion Formation in the Web," in *The Proceedings of the 12th Annual International Conference on Digital Government Research*, 2011, pp. 317–326.
7. J. Lule, *Understanding Media and Culture: An Introduction to Mass Communication*. 2017.
8. R. Pertierra, "Culture, Social Science, and the Conceptualization of the Philippine Nation State," *Kasarinlan: Philippine Journal Of Third World Studies*, vol. 12, no. 2. Kasarinlang: Philippine Journal of Third World Studies, pp. 5–24, 1996.
9. National Commission for Culture and the Arts, "NCCA Strategic Objectives," 2015.
10. X. Fan, J., Yun, L., Fei, D., & Di, "Relationships-aware Online Public Opinion Formation Model," *Int. Conf. Comput. Autom. Eng.*, pp. 196–199, 2010.
11. D. Murthy, *Twitter: Digital Media and Society Series*. Cambridge, UK: Polity Press, 2013.
12. M. Lash and K. Zhao, "Early Predictions of Movie Success: the Who, What, and When of Profitability," *J. Inf. Syst.*, 2016.
13. K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio, "Prediction of movies box office performance using social media," *Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. - ASONAM '13*, no. August, pp. 1209–1214, 2013.
14. D. Demir, O. Kapralova, and H. Lai, "Predicting IMDB movie ratings using Google Trends," no. 2, pp. 1–5, 2012.
15. C. Oh, Y. Roumani, J. K. Nwankpa, and H. F. Hu, "Beyond likes and tweets: Consumer engagement behavior and movie box office in social media," *Inf. Manag.*, vol. 54, no. 1, pp. 25–37, 2017.
16. H. Rui, Y. Liu, and A. Whinston, "Whose and what chatter matters? the effect of tweets on movie sales," *Decis. Support Syst.*, vol. 55, no. 4, pp. 863–870, 2013.
17. J. Gao, J. Otterbacher, and L. Hemphill, "Different Voices, Similar Perspectives? 'Useful' Reviews at the International Movie Database," *AMCIS 2013 Proc.*, no. May, pp. 1–12, 2013.
18. S. Mukhopadhyay, S. Conlon, and L. Simmons, "Consumer Feedback: Does Rating Reflect Reviewers' Feelings?," *AMCIS Proc.*, p. Paper 164, 2011.
19. D. Ray and M. Tarafdar, "How Does Twitter Influence a Social Movement?," *Res. Pap.*, vol. 2017, pp. 3123–3132, 2017.
20. J. Leavey, "Social media and public policy: what is the evidence?," *Alliance Useful Evid. Rep.*, no. September, p. 39, 2013.

21. J. R. Bautista and T. T. C. Lin, “Tweeting Social Support Messages After a Non-Celebrity’s Death: The Case of the Philippines’ #Fallen44,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 18, no. 11, pp. 641–646, 2015.
22. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 12, pp. 2825–2930, 2011
23. C. R. Soriano, M. D. G. Roldan, C. Cheng, and N. Oco, “Social media and civic engagement during calamities: the case of Twitter use during typhoon Yolanda,” *Philipp. Polit. Sci. J.*, vol. 37, no. 1, pp. 6–25, 2016.

Perception and Skill Learning for Augmented and Virtual Reality Learning Environments

Ng Giap Weng[1] and Angeline Lee Ling Sing[2]

[1]Knowledge Technology Research Unit, Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
[2] Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

nggiapweng@ums.edu.my, lingsing@hotmail.com

Abstract. This research modeled previously unarticulated experience of perception and skill learning in two different learning environments, namely Augmented Reality Learning Environment (ARLE) and Virtual Reality Learning Environment (ViRLE). An analytical literature review, primarily from Augmented Reality (AR) and Virtual Reality (VR) learning environments, human factor approach, user-based experimentation domains, methodological and contextual variations, was carried out to bring the dispersed research evidences together, organize and develop them into a meaningful conceptual structure. Sixty undergraduate participants, who were illiterate in computer subjects were selected based on participants' background. They provided their consent and actively participated as two equal groups of thirty participants, in both environments. This experiment was guided primarily by cognitive task analysis and user modelling techniques. Data elicited from this experiment included verbal protocols, video recording, observers' field notes, performance tests and responses to questionnaires. Participants' implicit mental models, such as Cognitive Model, Artefact Model and Task Model gradually evolved from numerous iterative cycles of re-construction, analyses and refinement from these data.

Keywords: Augmented Reality, Virtual Reality, Cognitive Task Analysis, User Modeling, Effectiveness, Usability.

1 Introduction

Augmented Reality (AR) techniques offer a potential solution to the problems associated with training learners to perform maintenance tasks. With AR, the computer offers added information in the learners' arena of view, usually in a Head-Mounted Display (HMD) worn by the learners which enriches or augments the learners' view of the real world [1]. AR is different from the Virtual Reality (VR), in which it effectively substitutes the real world maintenance environment by a simulated one [2]. AR can in practice provide the training direction and experience that could be provided in the virtual world or desktop environment, while permitting the learners to see and touch the physical objects. For example, AR may incorporate explained bolster for

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_38

naming framework parts or functionality of the system, or the presentation of documentation such as maintenance or manufacturing records. Furthermore, it would be possible for a remote expert to provide assistance by controlling the information presented by the system.

Learning technologies, loosely defined as technology-based tools used for purposes of education, have long been regarded as means of improving teaching and learning. In the 1980's, computers and videotape joined filmstrip, slide show, and microfiche systems were developed to teach English, mathematics, history and science to students from kindergarten to graduate school. Also in the 1980's began the surge of interest in interactive video and computer-aided instruction, and as a result, many institutions and organizations started to use and develop computer-based instructional materials. Sometime later interest turned towards the Internet and the World Wide Web (WWW) because the Web's massive record of data comprising of text, image, sound and video, and its ability to link information and people together, has captured the attention of educators who strive to offer inspiring learning environments to their students [3].

The concept of learning has changed around the world. The potential of AR and VR as learning environments was encouraging because these technologies offer learning through visualization on windows to learners. In addition, disability learners can access it through the windows without fear at any time and any place where the learners feel comfortable and convenience. In the learning context, learners can access and use the learning material appropriately to develop new knowledge, skills and attitudes to learn by making mistakes and/or being free of mistake without suffering the real consequences of errors. The learners have freedom to take risks, undertake adventurous or novel actions and explore the possible consequences. As such, AR and VR learning environments provide safe environment for learning to take place. Therefore, it is crucial for educators and learners to enhance their understanding of the issues, trends and opportunities associated with AR and VR continuously.

A research was conducted; focused on learners' perception on a learning task. Generally, this research created human cognitive assessment and learning scenarios, which allowed for precise control of complex stimulus representation. Computer accessory maintenance task was chosen because computers were ubiquitous in the university. Moreover, skill acquisition and inductive learning could be easier to identify and observable from novices than from the experts. The specific objectives of this research were a) to design AR and VR learning environments b) to outline how the theories could be applied to designing user interfaces which facilitate learning and use of AR and VR and finally c) to investigate whether the learners possessed the ability to be self-directed in adopting AR and VR as their learning tool and what is the level of their self-directed learning readiness.

2 Methodology

Sixty undergraduate participants provided their consent to participate in this experiment, and they provided some of their personal background history required by this research. Participants were divided into two equal groups (30 undergraduates participated in ARLE, and 30 undergraduates participated in ViRLE) and were briefed about their responsibilities before experimentations in Usability Laboratory. Their primary tasks were getting familiar with naming of related computer components (computer ports), knowing their related functions and understanding sequential steps for installing them. Video data, verbal protocol data with experimenter field notes were collected from these experiments for tracking, analyzing and interpreting task behavior and perception. In addition, all participants answered questionnaires and perform tests related to installation of computer components, as provided by experimenter.

2.1 ARLE and ViRLE Learning Features

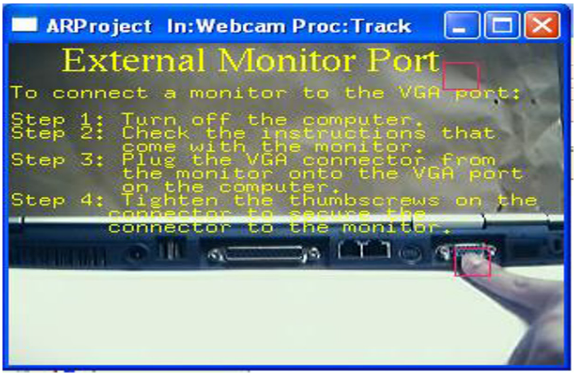
Two environments, namely ARLE and ViRLE systems (Figure 1 and Figure 2), were designed and developed in the Intelligent Visualisation Research Laboratory, School of Computing, Science, Engineering, University of Salford, United Kingdom using Visual C++ and OpenGL and runs under Windows NT and 2000 [4]. Virtual reality (VR) and augmented reality (AR) -- overlaying virtual objects onto the real world) offer interesting and wide spread possibilities to study different components of human behaviour and cognitive processes [5].

ARLE and ViRLE were designed to train the learners' basic function of computer ports in indoor setting and involved computer maintenance. They were designed for the learners who were keen to possess knowledge on basic functionality of computer hardware. It provided a typical task that described each computer port function and how to install the computer accessory in the correct procedures. These environments allowed the learners to learn the functionality of computer ports and sequential steps for installing computer accessory. The potential learners were likely to explore and visualize the function of computer hardware in the real and virtual environment, as well as practice and apply the task in a practical way.

In addition, the environments provided active interaction with the learning content, which could be revisited. The learners were allowed to perform learning through exploring and experiencing. ARLE and ViRLE allowed the learners to perform a self-directed learning for maintenance computer accessory maintenance. Moreover, the learning environments consider the aspect of transparent user interface; the learners could focus attention mainly on the learning contents.

ViRLE was developed based on VR solution involved at 3D objects like virtual rectangle boxes, arrows and words, which can be display on a computer screen. Learners in ViRLE could view the learning environments through windows and immerse in the setting. The purpose was to superimpose the instructions on the 'live' view of the computer ports with the boxes at the correct positions. A warning message was pop out to alert the learners to pull the computer accessory from computer port to avoid errors and mistakes.

ARLE was developed based on AR approach used image processing techniques to locate and track onto a key component in real time as the learners interact with the computer, and links to a small database contains relative positional information for the other components. As in ARLE, it tracked the computer ports, the annotations move with the computer ports as its positioning changes. The lines attaching the annotation tags with the computer ports follow the fitting visible mechanisms, permitting the learners to certainly detect the different parts as the view of the computer ports changes.



External Monitor Port

To connect a monitor to the VGA port:

Step 1: Turn off the computer.
Step 2: Check the instructions that come with the monitor.
Step 3: Plug the VGA connector from the monitor onto the VGA port on the computer.
Step 4: Tighten the thumbscrews on the connector to secure the connector to the monitor.

Fig. 1. ARLE System

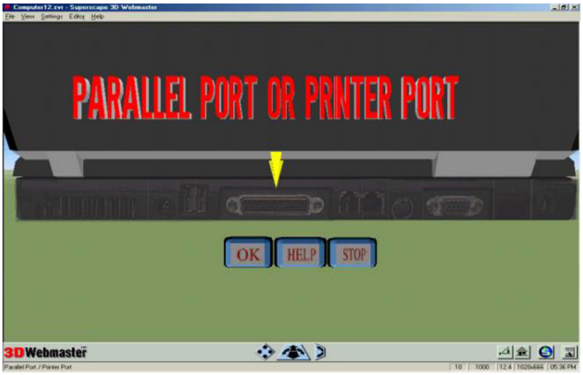

File View Settings Editor Help
PARALLEL PORT OR PRNTER PORT
OK
HELP
STOP
3DWebmaster
Parallel Port / Printer Port

Fig. 2. ViRLE System

3 Results and Discussions

An important result of this research, participants' implicit mental models such as Cognitive Model, Artefact Model and Task Model gradually evolved from numerous iterative cycles of re-construction, analyses and refinement from data collected from this experimentation.

Key Theme – 'Cognition'

Many researchers working in the domain of mixed reality have attempted to focus their research program on how information can effectively connect participants with their immediate environment by overlaying information, using especially annotations. Although it was known that information can be presented through various types of head mounted displays, and procedures for annotating information, such as descriptions of imperative features or instructions for execution physical tasks well established, there have been few reports which advances knowledge of participants' visual perception, especially participants in the augmented and virtual reality learning environment. In the present experiment, we examined participants' perception in mixed environment and empirically and successfully derived participants' mental model of maintenance task (perception). Worth cautioning was that this generalization was taken from a population of sixty participants, and a post-hoc research would have strengthen the argument for this model.

The user model derived from cognitive task analysis had detailed specification of the maintenance task to draw upon the knowledge from participants. Participants used their prior knowledge to achieve maintenance; although this was based on adequate maintenance approach and not maximizing their performance in term of using help facilities and creating maintenance value. The maintenance performance was controlled by their situation cognition and maintenance tools. The user model was suitable for integration of maintenance task and computerized maintenance system, which caused in designing and prototyping AR or VR systems. Moreover, the ARLE and ViRLE systems could avoid violating usability values and remove potential error which caused by poor system design. The designed ARLE and ViRLE also considered user-friendly interface, and user-centred principles to meet 'ergonomics' system design by taking human factors as centrality requirement. Furthermore, the system could customize the interactions among participants and the ARLE and ViRLE systems.

The method for arriving at this said participants' mental model was taken from sixty participants of the experiment, who were recruited randomly, with standardized experimental procedures. The procedures were generally guided by the work of [6] on human interacting with computers. Moreover, verbal protocol, video data, and field notes were taken during the studies. These data were further using task analyses techniques and user modelling techniques. In relation to the literature, this research was related broadly to the fruitful research program of the computer and the mind. It is John von Neumann who first constructed the computer and Alan Turing who subsequently compare the computer with the brain. Therefore, viewing from the perspective of the results, the methods, and the related literature, this research can be interpreted as having provided sufficient answers related generally to the participants 'perception.

Key Theme – 'Artefact'

This present research was conducted to extend the understanding of the usability, and its related performance of the augmented and virtual reality learning environment. The results of this research confirmed that both augmented and virtual reality meets good usability criteria. However, participants in augmented reality learning environment outperformed those in the virtual reality learning environment, and learning performance was affected by user-friendly learning environment.

ARLE was found to be a better instructional aid for computer accessory maintenance task than ViRLE. It could be seen that irrespective of the ARLE display type resulted in both lower assembly times and fewer assembly errors than the ViRLE. This was an important discovery because it served as upkeep for ARLE as a viable instructional aid for a maintenance task. It was speculated that the reduction in assembly time and errors were in part the result of having the maintenance instructions in the operator's direct visual workspace as opposed to just being near it. This reduced the number of head, eye, and body movements required by the operator to read the instructions, and aid to reduce the amount of information the operator needs to store in memory. Since the instructions were in the operators field of view while they were executing the task, they did not need to look at the instruction, turn to the assembly, remember what they just read, and perform the task. There was no need for the participants to have to turn and recheck the instructions since at the slight move of the eyes; the instructions were readily accessible using ARLE. Moreover, a remarkable result shown in the data was not only the assembly time better with the ARLE, but the standard deviations for participants' times were also reduced. This may be recognized to the fact that the ARLE environment offered a much more uniform instruction type than the ViRLE environment. For example, when using the ViRLE, how fast they were able to identify the 3D image, and how assured they were when they inserted a 3D element. This sometimes resulted in some participants double checking their work.

Worth mentioning is that the style of an interface, in terms of the shapes, fonts, colors and graphical elements that are used and the way they are combined, effects how the subjective level of pleasure to interact with. The more effective the use of imagery at the interface, the more engaging and entertaining it can be. Therefore, a good learning environment meets good usability criteria that led to better learning performance. Until recently, human-computer interaction has focused mainly on getting the usability right, with little thoughtfulness being paid to how to design aesthetically pleasing interfaces. Interestingly, current research proposes that the aesthetics of an interface can have positive effect on people's awareness of the system's usability. Moreover, when the 'look and feel' of an interface is pleasing (e.g., beautiful graphics, nice feel of the way the elements have been combined, ideal designed fonts, stylish use of images and colour) participants were likely to be more accepting of its usability (e.g., they may willing to wait a few seconds for a web site to download) [7]. Interactive design should not just be about usability per see, but should also contain aesthetic design, such as how pleasurable an interface is to look at (or listen to). The key is to get the right balance between usability and other design concerns, like aesthetics.

When comparing both augmented reality and virtual reality learning environment of this research, perhaps, it can be said that both systems are very attractive. Participants were able to go back to main environment and avoid themselves from lost in hyperspace when they were using the system. Basically, both systems are designed to train the participants who have or without knowledge computer accessories maintenance knowledge and skills. They could learn the computer accessory maintenance from ARLE and ViRLE systems. These systems can control the learning task done by participants such as giving instructions to prevent participants from wandering and lost in hyperspace. Both ARLE and ViRLE systems were also able to play sounds, which can make the learning more interesting. Although sound feedback can make the learning more interesting, however it was not for the evaluation in this research. The participants were able to feel that they were in the learning environment and the systems were able to retrieve information successfully. Unfortunately, there were some limitations with both systems. The display in ARLE system was not clear enough although this system used projector as a computer screen. Besides that, the participants could not view the ARLE learning environment in full screen mode in the real time because it would hang the system or distortion of the display on the screen if the participants were trying to change the mode of display. In addition to that, the participants found that the ARLE system was not accurate when using the pointing device to point at the computer port during their learning. This might delay and slow down the speed of learning process.

Key Theme – 'Task'

This research has attempted to understand how participants feel in the learning environment. In the ARLE, participants were in complete control of the learning task. They could decide what to learn and when to learn. They were able to avoid themselves from being lost in hyperspace when they are in the environment. The design of the interface and environment encourage participants to complete the tasks through the system. This system has contributed and played its role as an effective learning environment because it has the capability of representing computer accessory information. It was able to give participants memory retention and application after learning. The user interface design allowed participants to move easily and the environment helps them to have better understanding and visualizing of the computer accessory knowledge. In contrast, failure to visualize or understand the knowledge causes stress and loss of interest towards the computer accessory maintenance in ViRLE. The main limitation in this research was the cohort of participants were not sufficient to give a more significant statistically reliability.

Modelling a maintenance task with cognitive task analysis and user modelling techniques has led to suggest for better versions of the ARLE and ViRLE systems, especially the design of the user interface, which can be improved in term of it usability and to be more user-friendly. These techniques can attract those participants to learn more seriously by considering in term of their perspective and the way they thought. The interface of the system was designed to be transparent in order for the participants to focus directly on their task, instead of diverting their attention on correctness of understanding the functionalities of the system when using it. This was a task where system designers and participants have to cooperate in order to coordinate maintenance task between participants and the system itself. Participants could be a sufficient production in maintenance task; however, it might not be an efficient production without the ARLE and ViRLE systems. Hence, it was vital to recognize the interaction among participants and AR and VR system. It helped to solve participants' daily problems rather than delay their performance.

With this in mind, the AR and VR system could naturally use to enrich maintenance task and extend maintenance skill in task accomplishment especially for 'distance-based' jobs. The ARLE and ViRLE systems were designed to suit the participants in the sense of being usable, useful and learnable, because it was designed by shifting participants' prior knowledge to the system. The redesigned versions of ARLE and ViRLE systems needed to have a 'tailored environment' for participants to work effectively. Therefore, participants could accomplish within their professional capability on the ARLE and ViRLE systems in a safe, productive and healthy manner without feeling pressure. Furthermore, they were able to augment on their professional role by spending more time on non-routine maintenance task. It was essential to capture and review every subtask in maintenance process. To customize participants with their task, the ARLE and ViRLE environment must have the features of the everyday task environment of maintenance. The everyday task environment was important to system design, where explicit information would guide the participants to perform maintenance in adaptive environment. This research simplified the model of maintaining by integrating both automatic maintenance processing and controlled maintenance processing. The ARLE and ViRLE prototype that were designed were another AR and VR maintenance application. Further, this research had highlighted the significance of participants' participation in designing and prototyping ARLE and ViRLE systems.

Note also that the results from the think-aloud and observation protocols had one important implication. In general, the object's structure, object shape (what) and location (where), in particular, were crucial considerations if the images on the visual display unit (VDU) were to make sense. Object discrimination was part of learning set formation. The ultimate goal of designing ARLE and ViRLE was to provide participants with the possibilities of immersion and to act within learning environment. Given that all participants must ideally use their prior knowledge, expertise, values, beliefs and experience to visualize, select cues and signals, make sense of, interpret and interact with the objects in the learning environment, it was appropriate to include the participants' conceptual models in future designs of the learning environments.

Worth mention is this research confirmed and supported inductive reasoning [8]. It should not be forgotten, however, that teachers still have a significant role in mentoring, attention directing, and promoting participants' awareness, in the context of learning. Computer-aided-instructions and guidance systems like ARLE and ViRLE remained as important educational medium and helped facilities for learning. Other approaches towards learning process could include deductive learning for learning environments. Deriving from the previous paragraphs, it could be interpreted as this research had sufficiently addressed the learning as a process.

4 Recommendations and Future Direction

This research was intended to view learning process through the eyes of learners'. One of the most important topics worth discussing was the ways individual participant dealt with the issue of learning as a process, using various cognitive strategies and skills. Comparing among participants in this experiment, this research illuminated how viewing differently by each participating participants might mean for current and future intelligent/behavioral models of learning environments. Notably, this research had involved participants in significant ways: (a) the process of self-awareness, (b) confrontation, and (c) exposure to alternative conceptions of learning environment. The participants were moved from understanding their current learning perceptions and practices to planning future learning possibilities, especially with computer aided instructions. This research might effect a long-term change in the learning practice by allowing learners' the opportunity to become self-aware of their implicit beliefs, and direct examining of their learning processes. The results of this research suggest that a successful learning process was a joint product of learners' cognition and learning environment.

In the training and learning environment, learners can "generate a glowing passion amongst learners that is in marked contrast with the more common response of unwilling acceptance or outright hostility" [9]. Learning through exploration is considered as a good learning approach. It can be improved by expanding and including more navigation in the real world and/or adding more intelligent instructions in the system. In future, ARLE and ViRLE may be supplementing the existing learning environment if more aspects are taken into consideration such as psychology point of view and human factor approach. The feedback gained in the evaluation shows that learners prefer to perform an interactive self-directed learning activity. A part from expanding this system, learners who were physically enabled can access the information at anywhere and anytime.

Future researchers could increase number of respondents in order to gain more accurate results and get more suggestions about how to improve the system. They can make some improvement on the system so that it becomes more user-friendly system. It was hopefully that with this recommendation, at future researchers can develop more effective learning system for maintenance. Future research should also clarify further attention mechanisms through which we select and control, what we see and hear, learn and remember and lastly think and do.

5 Conclusion

Based on the results presented, AR was found to be a better instructional aid for computer accessory maintenance task than VR. It can be seen that irrespective of the AR display resulted in both lower maintenance times and less maintenance mistakes than the VR. This is an imperative finding because it serves as support for AR as a practical instructional aid for a maintenance task.

It was speculated that the lessening in assembly time and errors is in part the result of having the assembly instructions in the operator's direct visual workspace as opposed to just being near it. This reduces the number of head, eye, and body movements required by the operator to read the commands, and may also help lessen the amount of information the operator needs to store in memory. Since the commands are in the operators field of view while they are carrying out the task, they don't need to look at the instruction, turn to the assembly, remember what they just read, and perform the task. There is no need for the operator to have to turn and recheck the instructions since at the slight move of the eye, the instructions are readily accessible using AR.

The potential of using AR techniques for computer accessory maintenance has been discussed. From the work that has taken place so far, the anticipated benefits of this technology are numerous. The learners can be guided through the various maintenance steps interactively, working at their-own pace, but in the real environment. Training through direct involvement is reflected to be more effective than training through facts of information, and the AR approach supports this analysis. The technology offers a simple way of progressing to the maintenance of more complex equipment. This approach is promoting 'active' training, both in psychological and physical sense, and will inspire the learners to have various thinking perceptions, which should prepare them better for their other day-to-day activities.

References

1. Janin, A.L., Mizell, D.W., & Caudell, T.P. (1993). Calibration of Head-Mounted Displays for Augmented Reality Applications. Proceedings of IEEE VRAIS '93, pp. 246-255. IEEE Press. Available: http://www.hitl.Washington.edu/ scivw/scivw-ftp/citations/Augmented-Reality-list, last accessed 2018/02/21.
2. Azuma, R. (1997). A Survey of Augmented Reality, In Presence: Teleperators and Virtual Environments. Available: http://www.cs.unc.edu/~azuma/AR presence.pdf, last accessed 2018/02/25.
3. Jonassen, D.H., Peck K.L., & Wilson, B.G. Learning With Technology: A Constructivist Perspective. New Jersey: Prentice-Hall, Inc (2006).
4. Garvey, D. A Software Framework for Augmented Reality Applications. MPhil Dissertation. University of Salford: Salford, UK (2005).
5. Dunser, A., Steinbugl, K., Kaufmann, H., & Gluck, J. (2006) Virtual and Augmented Reality as Spatial Ability Training Tools. Retrieved from https://dl.acm.org/citation.cfm?id=1152776
6. Card, S.K., Moran, T.P. & Newell, A. The Psychology of Human-Computer Interaction. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc, Publishers (1983).
7. Preece, J., Roger, Y., & Sharp, H. Interaction Design: Beyond Human-Computer Interaction. John Wiley & Son: USA (2015).
8. Haverty, L.A., Koedinger, K.R., & Klahr, D. Solving Inductive Reasoning Problems in Mathematics. In Greeno, J.G. (Executive Ed.), Cognitive Science: A Multidisciplinary Journal, 24 (2), pp. 249-298. Elsevier Science: USA (2000).
9. Shneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interface, 6th Edition. Addison-Wesley: Reading, MA (2017).

Application of Newton-4EGSOR Iteration for Solving Large Scale Unconstrained Optimization Problems with a Tridiagonal Hessian Matrix

Khadizah Ghazali[1], Jumat Sulaiman[1], Yosza Dasril[2], and Darmesah Gabda[1]

[1] Mathematics with Economics Programme, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia.
[2] Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia.

Corresponding addresses
khadizah@ums.edu.my, jumat@ums.edu.my, yosza@utem.edu.my, darmesah@ums.edu.my

Abstract. Solving the unconstrained optimization problems using Newton method will lead to the need to solve linear system. Further, the Explicit Group iteration is one of the numerical methods that has an advantage of the efficient block iterative method for solving any linear system. Thus, in this paper to reduce the cost of solving large linear system, we proposed a combination between Newton method with four-point Explicit Group (4-point EG) block iterative method for solving large scale unconstrained optimization problems where the Hessian of the Newton direction is tridiagonal matrices. For the purpose of comparison, we used combination of Newton method with basic iterative method namely successive-over relaxation (SOR) point iteration and Newton method with two-point Explicit Group (2-point EG) block iterative method as reference method. The proposed method shows that the numerical results were more superior compared to the reference methods in term of execution time and number of iteration.

Keywords: Explicit Group iteration, Newton method, Unconstrained optimization problems.

1 Introduction

The model of unconstrained optimization problems arise in a variety of situations, with most commonly when the problem formulation is simple. More complex formulations often involve explicit well designed constrains. Still, many problems with constraints are frequently converted to unconstrained problems as in [1-3]. Unconstrained optimization can be solved by using direct search methods [4-6] and gradient descent methods [7-14]. The direct search method do not need differentiability or even continuity of the objective function, but it display slower convergence rates than

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_39

gradient descent methods [6] and tend to be more reliable for problems on noisy functions [15]. Numerous researchers used the gradient descent method in varying forms such as steepest descent method [7], Newton's method [8], modified Newton method [9], Levenberg-Marquardt's method [10], conjugate gradient method [11], Quasi-Newton method [12], Boryden-Flecther-Goldfarb-Shanno (BFGS) method [13] and Powell's method [14]. Generally, unconstrained problems covers basic properties of solutions and algorithms. The most useful algorithms are listed in [15], where it also stated that the partitioned quasi-Newton method, the limited memory BFGS method and Newton method were suitable for large scale optimization. Thus, in this paper, only Newton method will be discussed since in theoretically it is the fastest unconstrained optimization method [16]. According to [17] unconstrained optimization problems are considered large scale cases, where the dimensions of the problems are up to 10^3. Problems involving large scale unconstrained optimization have been discussed in [16-20].

Basically, Newton method is one of the most popular methods due to its attractive quadratic convergence [16-20]. Moré and Sorensen [16] presented and explored on Newton's method for unconstrained minimization for large scale problem and they pointed out that it is possible to reduce amount of work and storage for the Newton method by using the Cholesky decomposition of symmetric matrix. In that regard, Gundersen and Steihaug [18] also shown that the ratio of the number of arithmetic operation of Newton's method is constant per iteration for a large class of sparse matrices.

Newton methods also depends on the choosing starting point and sometimes the storage for calculating the inverse of the Hessian could be very expensive [19]. In order to avoid the difficulties, Sorensen [21], Sisser [22], Dasril et. al. [23] and Gill and Murray [24] have modified the Newton iteration. Thus, in this paper, we proposed new algorithm based on the idea of combining the Newton method with 4-point EG iterative methods, namely Newton-4EG for solving a large scale unconstrained optimization problems to overcome this difficulties. This combination is motivated by the advantage of the EG iterative method which is known as one of the efficient block iterative methods which had been demonstrated by Evan [25], Yousif and Evans [26,27], Abdullah [28] and Othman and Abdullah [29,30]. Although previous researchers have discussed Newton method and EG iterative methods extensively, but combining them for solving large scale unconstrained optimization is a new approach.

To investigate the capability of Newton-EG method, let us consider a large scale unconstrained optimization problem formulated as

$$\min_{\underline{x}\in\mathbb{R}^n} f(\underline{x}) \tag{1}$$

where the objective function $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. In the process of finding the minimum value on the problem (1) using the Newton method, a search direction name as Newton direction is required. This Newton direction can be obtained by solving the linear system resulting from problem (1) and the computation of it could be time consuming since we are dealing with large scale problem. Thus, in this paper, we approximate the Newton direction by using 4-point EG block iterative method as an inner iteration and find an approximate solution for problems (1) by using Newton method as an outer iteration. For the purpose of comparison, we considered a combination of Newton method with SOR point iterative method, Newton-SOR and a combination of newton method with 2-point EG block iterative method, Newton-2EGSOR.

2 The Formulation of Newton Scheme with Tridiagonal Hessian Matrix

In this section, we start by approximate the objective function $f(x)$ in problem (1) around the current point $\underline{x}^{(k)}$ through the first three terms of Taylor series expansion used in unconstrained optimization;

$$f(\underline{x}) \approx f(\underline{x}^{(k)}) + \left[\nabla \underline{f}(\underline{x}^{(k)})\right]^T (\underline{x} - \underline{x}^{(k)}) + \frac{1}{2}(\underline{x} - \underline{x}^{(k)})^T \nabla^2 f(\underline{x}^{(k)})(\underline{x} - \underline{x}^{(k)}), \tag{2}$$

where $\nabla \underline{f}(\underline{x}^{(k)})$ represent the gradient of $f(\underline{x})$ and $\nabla^2 f(\underline{x}^{(k)}) = \mathbf{H}(\underline{x}^{(k)})$ denote as the Hessian matrix of $f(\underline{x})$.

Notices that, the approximation (2) is in the form of quadratic function, hence the right-hand side of (2) minimized at

$$\underline{x} = \underline{x}^{(k)} - \left[\mathbf{H}(\underline{x}^{(k)})\right]^{-1} \nabla \underline{f}(\underline{x}^{(k)}). \tag{3}$$

Since at minimum of $f(x)$, its gradient vector is zero and $\mathbf{H}(\underline{x}^{(k)})$ is symmetric because $f(x)$ is twice continuous differentiable. Therefore, equation (3) obtained by differentiating it with respect to $\underline{x}$ and equating the resulting expression to zero.

The next Newton iteration is prepared by updating $\underline{x}$ as $\underline{x}^{(k+1)}$. Calling $\underline{d}^{(k)} = \underline{x}^{(k+1)} - \underline{x}^{(k)}$ then the search direction is obtained by solving

$$\mathbf{H}(\underline{x}^{(k)})\underline{d}^{(k)} = -\nabla \underline{f}(\underline{x}^{(k)}), \tag{4}$$

whose solution can be written as

$$\underline{d}^{(k)} = -\left[\mathbf{H}(\underline{x}^{(k)})\right]^{-1} \nabla \underline{f}(\underline{x}^{(k)}). \tag{5}$$

This search direction (5) is a descent direction (commonly called the Newton direction) since it satisfies

$$\left[\nabla \underline{f}(\underline{x}^{(k)})\right]^T \underline{d}^{(k)} = -\left[\nabla \underline{f}(\underline{x}^{(k)})\right]^T \left[\mathbf{H}(\underline{x}^{(k)})\right]^{-1} \nabla \underline{f}(\underline{x}^{(k)}) < 0, \tag{6}$$

if $\mathbf{H}(\underline{x}^{(k)})$ is positive definite.

Alternatively, we can said that when the iteration are sufficiently close to the minimum, $\underline{x}^{(*)}$, this Newton scheme have quadratic convergence if for large k:

$$\left\|\underline{x}^{(k)} - \underline{x}^{(*)}\right\| \rightarrow C\left\|\underline{x}^{(k-1)} - \underline{x}^{(*)}\right\|^2 \rightarrow 0 \tag{7}$$

where C is a positive constant.

2.1 Tridiagonal Hessian Matrix

We shall consider a tridiagonal matrix $\mathbf{H}(\underline{x}^{(k)})$ in the form [31];

$$\mathbf{H}(\underline{x}^{(k)}) = [h_{i,j}] = \begin{bmatrix} b_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & \cdots & 0 \\ 0 & a_3 & b_3 & c_3 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & a_n & b_n \end{bmatrix}, \tag{8}$$

with $h_{i,i} > 0 \; \forall \; i$, $h_{i,j} < 0$ for $i \neq j$, $h_{i,j} = 0 \; \forall \; |i-j| > 1$ and $i, j = 1, 2, \dots, n$. Since matrix (8) is symmetric, thus we have

$$a_i = c_{i-1}, i = 2, 3, \dots, n. \tag{9}$$

Furthermore, this tridiagonal matrix $\mathbf{H}(\underline{x}^{(k)})$ has a diagonal entries that satisfies the inequality

$$|b_i| \geq |a_i| + |c_i|, \forall \; i = 01,2, \dots, n \tag{10}$$

where $a_1 = c_n = 0$, so that the condition of positive definite can be fulfilled.

3 Derivation of proposed Iterative Methods

Solving problem (1) with Newton method in its basic form will need us to find the inverse of Hessian matrix $\mathbf{H}(\underline{x}^{(k)})$ as stated in equation (5). Since the coefficient matrix $\mathbf{H}(\underline{x}^{(k)})$ is a large sparse matrix, choosing direct method such as Gauss elimination method or simultaneous method will involve tedious work, which leads to a long computational work. Thus, we proposed method by using iterative method as in [32,33]. Since equation (4) is known as a linear system, let the linear system be re-written as

$$\mathbf{A}\underline{d} = \underline{b} \tag{11}$$

where,

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{2,1} & a_{3,1} & \dots & a_{1,n} \\ a_{1,2} & a_{2,2} & a_{3,2} & \cdots & a_{3,n} \\ a_{1,3} & a_{2,3} & a_{3,3} & \cdots & a_{4,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & a_{3,n} & \cdots & a_{n,n} \end{bmatrix}, \underline{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n \end{bmatrix} \text{ and }, \underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

To evaluate the performance of the 4-point Newton-EG iterative method for solving the linear system (11), the Newton-SOR and 2-point Newton-EG iterative methods used as the reference methods. The following subsections will discuss on the formulation of the SOR and 4-point EG iterative methods.

3.1 Formulation of SOR Iterative Method

To derive the formulation of SOR iterative method, let the real coefficient matrix **A** of the linear system (11) be decomposed as summation of three matrixes as follows;

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U} \tag{12}$$

in which **D** is the nonzero diagonal entries of **A**, **L** is strictly lower matrix and **U** is strictly upper matrix. Then, the general form of the SOR iterative method for solving the linear system (11) can be stated in vector form as [34,35]

$$\underline{d}^{(k+1)} = (\mathbf{D} - \omega\mathbf{L})^{-1}(\omega\mathbf{U} - (\mathbf{1} - \omega)\mathbf{D})\underline{d}^{(k)} + \omega(\mathbf{D} - \omega\mathbf{L})^{-\mathbf{1}}\underline{b} \tag{13}$$

The formulation of SOR iterative method with equation (13) can be represented as the i^{th} component of the vector $\underline{d}$;

$$d_i^{(k+1)} = (1 - \omega)d_{i-1}^{(k+1)} + \frac{\omega}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}d_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}d_j^{(k)}\right) \tag{14}$$

where ω represent as a relaxation factor with the optimal value in the range of $[1,2)$. Obviously, the formulation in equation (14) can be categorized as one of the point iterative methods. Based on the concept of this point SOR iterative method, the next subsection will discuss the formulation of block SOR method, known as the Explicit Group (EG) iterative method.

3.2 Formulation of Explicit Group (EG) Iterative Method

Apart from the concept of point iterative methods, Evans [25] has proposed four point block iterative methods via the Explicit Group iterative method for solving large linear systems. This study was continued extensively in [26-30] for verifying the effectiveness of block iterative methods. Due to the advantages of block iterations, this paper is to examine the performance of 4-point Newton-EG iterative method for solving a large linear system generated by imposing the Newton method to large scale unconstrained optimization problem (1). Thus, this section shows on how to derive the formulation of 4-point EG iterative method by using the same steps as 2-point EG approach. To derive the formulation of the proposed iterative method, let us consider a group of two points from the linear system (11) as follows [36,37];

$$\begin{bmatrix} a_{i,i} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{bmatrix} \begin{bmatrix} d_i \\ d_{i+1} \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{15}$$

where,

$$S_t = b_{i+t-1} - \sum_{j=1}^{i-1} a_{i+t-1,j}d_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i+t-2,j}d_j^{(k)},\ t = 1,2.$$

From equation (15) the general form of the 2-point EG iteration can be stated as;

$$d_i^{(k+1)} = (1-\omega)d_{i-1}^{(k+1)} + \omega\left(\frac{a_{i+1,i+1}S_1 - a_{i,i+1}S_2}{\alpha}\right), \tag{16}$$

$$d_{i+1}^{(k+1)} = (1-\omega)d_i^{(k+1)} + \omega\left(\frac{a_{i,i}S_2 - a_{i+1,i}S_1}{\alpha}\right), \tag{17}$$

where $\alpha = (a_{i,i})(a_{i+1,i+1}) - (-a_{i,i+1})(-a_{i+1,i})$. To implement the 4-point EG iterative method, a group of four points from the linear system (11) is considered as [36,37];

$$\mathbf{G}\underline{u} = \underline{s} \tag{18}$$

where,

$$\mathbf{G} = \begin{bmatrix} a_{i,i} & a_{i,i+1} & a_{i,i+2} & a_{i,i+3} \\ a_{i+1,i} & a_{i+1,i+1} & a_{i+1,i+2} & a_{i+1,i+3} \\ a_{i+2,i} & a_{i+2,i+1} & a_{i+2,i+2} & a_{i+2,i+3} \\ a_{i+3,i} & a_{i+3,i+1} & a_{i+3,i+2} & a_{i+3,i+3} \end{bmatrix}, \underline{u} = \begin{bmatrix} d_i \\ d_{i+1} \\ d_{i+2} \\ d_{i+3} \end{bmatrix} \text{ and } \underline{s} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}.$$

Similarly, following the same steps in equation (15), thus the 4-point EGSOR iterative method can be easily stated as

$$\underline{u}^{(\boldsymbol{k+1})} = (1-\omega)\underline{u}^{(\boldsymbol{k})} + \omega\mathbf{G}^{-1}\underline{s} \tag{19}$$

Therefore, by using equation (11) and (19), we proposed the algorithm of 4-point Newton-EGSOR iterative method for solving problem (1), as stated in Algorithm 1.

Algorithm 1. Newton-4EGSOR Scheme

```
i.   Initialize
         Set up the objective function : $f(\underline{x})$, $f(\underline{x}^*) \leftarrow \mathbb{R}$, $\underline{x}^{(\mathbf{0})} \leftarrow \mathbb{R}^n$, $\varepsilon_1 \longleftarrow 10^{-6}$,
         $\varepsilon_2 \longleftarrow 10^{-8}$, $n \longleftarrow \{1000, 5000, 10000, 20000, 30000\}$.
ii.  For $j = 1,2,\ldots,n$, implement
     a.  Set $\underline{d}^{(0)} \longleftarrow 0$
     b.  Calculate $f(\underline{x}^{(k)})$
     c.  For $i = 1,2,\ldots,n$, calculate iteratively equation (11) by using equation
         (19).
     d.  Check the convergence test, $\|\underline{d}^{(k+1)} - \underline{d}^{(k)}\| < \varepsilon_2$. If yes, go to step (e).
         Otherwise go back to step (b)
     e.  For $i = 1,2,\ldots,n$, calculate $\underline{x}^{(k+1)} \longleftarrow \underline{x}^{(k)} + \underline{d}^{(k)}$
     f.  Check the convergence test, $\|\nabla \underline{f}(\underline{x}^{(k)})\| \leq \varepsilon_1$. If yes, go to (iii). Other-
         wise go back to step (a)
iii. Display approximate solutions.
```

4 Numerical Experiments

The Algorithm 1 was tested on three test functions taken from [38] and [39] and implemented in C language. For each test function we considered five numerical experiments with dimensions varied from 1000 to 30000 variables as listed in Table 1 and tested with three different initial points, $\underline{x}^{(\mathbf{0})}$ which is randomly selected but closer to the solution point, $\underline{x}^*$. In our numerical experiments, Algorithms 1 is stopped when $\|\nabla f(\underline{x})\| < 10^{-8}$ for inner iteration and $\|\nabla f(\underline{x})\| < 10^{-6}$ for outer iteration. For proper comparison, we used combination Newton method with the same subroutines. Thus, we compare the efficiency of our proposed method with Newton-2GSSOR iteration and Newton-SOR iteration. The efficiency of our proposed methods are evaluated based on the comparison of number of inner iterations, of number of outer iterations and the execution time. The details of the three test functions are given as follows;

Example 1: Generalized Tridiagonal 1 Function [38]

$$f(\underline{x}) = \sum_{i=1}^{n}(x_i + x_{i+1} - 3)^2 + (x_i - x_{i+1} + 1)^4 \tag{20}$$

This function has a global minimum, $f^* = 0$ at $x^* = (1,2)$ with Fig. 1.(a) shows the graph of this function when $n = 2$. The used starting points, $\underline{x}^{(\mathbf{0})}$ were:

(a) $\underline{x}^{(\mathbf{0})} = (2.0,2.0,\dots,2.0,2.0)$
(b) $\underline{x}^{(\mathbf{0})} = (0.0,2.0,\dots,0.0,2.0)$
(c) $\underline{x}^{(\mathbf{0})} = (1.0,2.0,\dots,1.0,2.0)$

Example 2: NONSCOMP Function [38]

$$f(\underline{x}) = (x_1 - 1)^2 + \sum_{i=2}^{n} 4\left(x_i - x_{i-1}^2\right)^2 \tag{21}$$

This function has a global minimum, $f^* = 0$ at $x_i^* = 1$, for $i = 1,2,\dots,n$ with Fig. 1.(b) shows the graph of this function when $n = 2$. The used starting points, $\underline{x}^{(\mathbf{0})}$ were:

(a) $\underline{x}^{(\mathbf{0})} = (3.0,3.0,\dots,3.0,3.0)$
(b) $\underline{x}^{(\mathbf{0})} = (1.5,1.5,\dots,1.5,1.5)$
(c) $\underline{x}^{(\mathbf{0})} = (1.5,1.0,\dots,1.5,1.0)$

Example 3: Dixon and Price Function [39]

$$f(\underline{x}) = (x_1 - 1)^2 + \sum_{i=2}^{n} i(2x_1^2 - x_{i-1})^2 \tag{22}$$

This function has a global minimum, $f^* = 0$ at $x_i^* = 2^{-\left[\frac{2^i-2}{2^i}\right]}$, for $i = 1,2,\dots,n$ with Fig. 1.(c) shows the graph of this function when $n = 2$. The used starting points, $\underline{x}^{(\mathbf{0})}$ were:

(a) $\underline{x}^{(\mathbf{0})} = (1.0,1.0,\dots,1.0,1.0)$
(b) $\underline{x}^{(\mathbf{0})} = (0.6,0.6,\dots,0.6,0.6)$
(c) $\underline{x}^{(\mathbf{0})} = (0.6,1.0,\dots,0.6,1.0)$

Even though, our starting points were selected randomly, but whenever a starting point was given in the literature, we used it as option (a). The efficiency comparison results for the execution time (seconds) and the number of iteration are tabulated in Table 1. With respect to the graph of functions in Fig. 1, it is obvious that the 3D plots showing the existence of the local minimum with an optimum value f^* at optimum point $\underline{x}^*$ such that problem (1) can be solved. From Table 1, it is apparent that the number of inner iteration for our proposed method decreased from the referenced method.

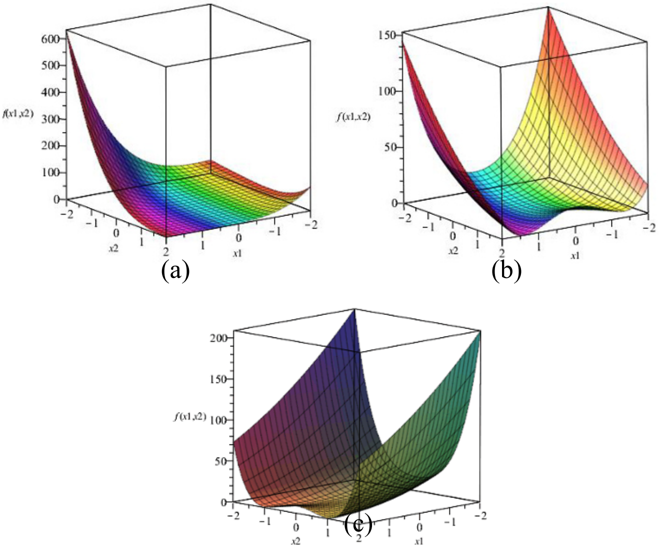


Fig. 1. (a) Generalized Tridiagonal 1 function in 3D, (b) NONSCOMP function in 3D and (c) Dixon and Price function in 3D

5 Conclusion

As shown in this paper, the combination of Newton method with 4EGSOR iterative method has speed up the process for solving large scale unconstrained optimization problems with a tridiagonal Hessian matrix. This can be seen through the execution time and the number of iteration produced as a result of our implementation proposed algorithm. According to Table 1, the numerical results show that our proposed algorithm reduced the number of inner iteration better than the reference algorithm with less execution time (second). There is approximately 23.33-99.68% and 7.69-89.88% reduction in total execution time for our proposed method compare to Newton-SOR and Newton-2EGSOR. Thus, the numerical comparison illustrates that our proposed method is potentially efficient for solving large scale unconstrained optimization problems.

Acknowledgement. The authors are grateful for the fund received from Universiti Malaysia Sabah upon publication of this paper (GUG0160-2/2017).

References

1. Laptin, Y. P.: An Approach to The Solution of Nonlinear Unconstrained Optimization Problems. Cybernetics and systems Analysis, **45**(3), 497-502 (2009).
2. Simon, D. and Tien, L. C.: An efficient method for unconstrained optimization problems of nonlinear large mesh-interconnected systems IEEE Transaction on Aerospace and Electronic Systems **38**(1), 128-136 (2002)
3. Jonathan, H. M.: Optimization Algorithms Exploiting Unitary Constraints. IEEE Transaction on Signal Processing **50**(3), 635-650 (2002)
4. Nelder, J. A. and Mead, R.: A simplex methos for function minimization, The Computer Journal **7,** 308-313 (1965)
5. Audet, C. and Dennis, Jr. J. E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(2), 188-217 (2006)
6. Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E.: Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM J. Optim. **9**, 112-147 (1998)
7. Shi, Z-J. and Shen, J.: Step-size Estimation for Unconstrained Optimization Methods *Comput. Appl. Math.* **24**(3), 399-416 (2005)
8. Babaie-Kafaki, S.: Emrouznejad A. (eds) Big Data Optimization: Recent Developments and Challenges. Studies in Big Data, vol 18. Springer DOI 10.1007/978-3-319-30265-2_17 (2016)
9. Kaniel, S. and Dax, A.: A modified Newton's method for unconsfrained minimization *SIAM J. Numerical Analysis* **16**(2), 324-331 (1979)
10. Higham, D. J.: Trust Region Algorithms And Timestep Selection *SIAM J. Numer. Anal.* **37**(1), 194-210 (1999)
11. Andrei, N.: An adaptive conjugate gradient algorithm for large-scale unconstrained optimization Journal of Computational and Applied Mathematics **292,** 83-91 (2016)
12. Aderibigbe, F. M., Adebayo, K. J. and Dele-Rotimi, A. O.: On Quasi-Newton Method for Solving Unconstrained Optimization Problems *America Journal of Applied Mathematics* **3**(2), 47-50 (2015)
13. Liu, D. C. and Nocedal, J.: On The Limited Memory Bfgs Method For Large Scale Optimization *Mathematical Programming* **48,** 503-528 (1989)
14. Powell, M. J. D.: Numerical Analysis (Dundee, 1983), Lecture Notes in Mathematics, vol. **1066,** 122–141 Springer, Berlin (1984)
15. Nocedal, J.: Theory of Algorithms for Unconstrained Optimization. Acta Numerica **1,** 199-242 (1992)
16. Moré, J. and Sorensen, D.: Newton's method, In:*Studies in Numerical Analysis, Golug, G. (ed)* (Washington DC: The Math. Association of America) 29-82 (1984)
17. Roma, M.: Large Scale Unconstrained Opti. Encyclopedia of Opti. Springer, United States (2001)
18. Gundersen, G, and Stiehaug, T.: On large-scale unconstrained optimization problems an higher order methods Optimization Methods and Software **25**(3). 337-58 (2010)
19. Sun, W. and Yuan, Y.: Optimization Theory and Methods-Nonlinear Prog. Springer, United States (2006)
20. Nocedal, J. and Wright, S. J.: Numerical Optimization 2nd end. Springer-Verlag, Berlin (2000)
21. Sorensen, D.: Newton's Method with a Model Trust Region Modification SIAM Journal on Numerical Analysis **19**(2) 409-26 (1982)
22. Sisser, F. S.: A modified Newton's method for minimization. J of Opt. and Application **38**(4) 461-82 (1982)

23. Dasril, Y., Mohd, I. and Mamat, M.: Proc. Int. Conf. on Integrating Technology. In The Matematical Sciences(2004), USM, Malaysia 426-32 (2004)
24. Gill, P. E. and Murray, W.: Newton-Type Methods for Unconstrained and Linearly Constrained Optimization. Mathematical Programming **7**(1) 311-50 (1974)
25. Evans, D. J.: Group explicit iterative methods *Int. J. Computer Maths.* **17** 81-108 (1985)
26. Yousif, W. S. and Evans, D. J.: Explicit group over-relaxation methods for solving elliptic partial differential equations *Mathematics and Computer in Simulations* **28** 453-66 (1986)
27. Yousif, .W. S. and Evans, D. J.: Explicit de-coupled group iterative methods and their implementations, Parallel Algorithms and Applications **7.** 53-71 (1995)
28. Abdullah, A. R.: 1991 The four point explicit decoupled group (EDG) method: A fast Poisson solver. *Int. J. Computer Maths*. **38.** 61-70 (1991)
29. Othman, M. and Abdullah, A. R.: An efficient four points modified explicit group poisson solver *Intern. J. of Computer Maths.* **76.** 203-17 (2000)
30. Othman, M., Abdullah, A. R. and Evans, D. J.: A parallel four point modified explicit group iterative algorithm on shared memory multiprocessors, Parallel Algorithms and Applications **19**(1). 705-717 (1972)
31. Li, H-B., Huang, T-z, Lui, X-P and Li, H.: On the inverses of general tridiagonal matrices. Linear Algebra and its Applications **433**. 965-983 (2010)
32. Sulaiman, J., Hasan, M. K., Othman, M. and Karim, S. A. A.: Fourth-order solutions of nonlinear two-point boundary value problems by Newton-HSSOR iteration *AIP Conference Proceedings 1602*, 69-75 (2014)
33. Sulaiman, J., Hasan, M. K., Othman, M. and Karim, S. A. A.: Application of Block Iterative Methods with Newton Scheme for Fisher's Equation by Using Implicit Finite Difference *Jurnal Kalam.* **8**(1). 039-46 (2015)
34. Young, D. M.: Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc. **76**. 92-111 (1954)
35. Young, D. M.: Iterative solution of large linear system, Academic Press, London (1971)
36. Sulaiman, J., Hasan, M. K., Othman, M. and Karim, S. A. A.: Newton-EGMSOR Methods for Solution of Second Order Two-Point *Journal of Mathematics and System Science* **2** 185-90 (2012)
37. Sulaiman, J., Hasan, M. K., Othman, M. and Karim, S. A. A.: Numerical solutions of nonlinear second-order two-point boundary value problems using half-sweep SOR with Newton method. *Journal of Concrete & Applicable Mathematics*. **11**(1), 112-20 (2013)
38. Andrei, N.: An unconstrained optimization test function collection, Advanced Modeling and Optimization, **10**(1): 147-161 (2008)
39. Laguna, M. and Marti, R.: Experimental Testing of Advanced Scatter Search Designs for Global Optimization of Multimodal Functions *Journal of Global Optimization* **33**(2), 235-55 (2005)

Table 1. Comparison of number of iteration and execution time (second) for Newton-SOR, Newton-2EGSOR and Newton-4EGSOR method.

	Example	$\underline{x}^{(0)}$	Method	Order of Hessian matrix, *n*					Total exe-cution time
				1000	5000	10000	20000	30000	
No. of inner iteration (No. of outer iteration) / Execution time	1	(a)	Newton-SOR	66(4)/0.01	66(4) / 0.04	66(4) / 0.08	66(4) / 0.11	66(4) / 0.15	0.39
			Newton-2EGSOR	38 (4) / 0.01	38 (4) / 0.03	38 (4) / 0.07	38 (4) / 0.09	38 (4) / 0.13	0.33
			Newton-4EGSOR	25(4) / 0.00	25(4) / 0.02	25(4) / 0.06	25(4) / 0.07	25(4) / 0.10	0.25
		(b)	Newton-SOR	237(8) / 0.01	237(8) / 0.04	237(8) / 0.10	237(8) / 0.15	237(8) / 0.23	0.53
			Newton-2EGSOR	183(8) / 0.01	183(8) / 0.04	183(8) / 0.06	183(8) / 0.11	183(8) / 0.16	0.38
			Newton-4EGSOR	178(8) / 0.01	178(8) / 0.04	178(8) / 0.06	178(8) / 0.11	178(8) / 0.16	0.38
		(c)	Newton-SOR	124(5) / 0.01	124(5) / 0.03	124(5) / 0.05	124(5) / 0.08	124(5) / 0.13	0.30
			Newton-2EGSOR	101(5) / 0.00	101(5) / 0.02	101(5) / 0.04	101(5) / 0.07	101(5) / 0.10	0.23
			Newton-4EGSOR	96(5) / 0.00	96(5) / 0.02	96(5) / 0.04	96(5) / 0.07	96(5) / 0.10	0.23
	2	(a)	Newton-SOR	76137(322) / 1.79	86359(337) / 10.02	90525(202) / 20.86	90525(202) / 41.77	90525(202) / 62.40	136.84
			Newton-2EGSOR	6557(8) / 0.07	6557(8) / 0.33	6557(8) / 0.67	6565(15) / 1.34	6565(15) / 1.94	4.35
			Newton-4EGSOR	213(8) / 0.01	213(8) / 0.05	213(8) / 0.10	213(8) / 0.18	213(8) / 0.10	0.44
		(b)	Newton-SOR	8523(6) / 0.20	8632(8) / 1.00	8632(6) / 1.99	8632(8) / 3.87	8632(8) / 5.92	12.98
			Newton-2EGSOR	1920(6) / 0.04	1920(6) / 0.19	1920(6) / 0.39	1932(8) / 0.82	1932(8) / 1.21	2.65
			Newton-4EGSOR	173(6) / 0.00	173(6) / 0.04	173(6) / 0.08	173(6) / 0.15	173(6) / 0.21	0.48
		(c)	Newton-SOR	8437(6) / 0.20	8437(6) / 1.00	8437(6) / 2.16	8437(6) / 4.34	8437(6) / 6.46	14.16
			Newton-2EGSOR	1696(6) / 0.07	1696(6) / 0.27	1696(6) / 0.54	1696(6) / 094	1696(6) / 1.32	3.14
			Newton-4EGSOR	143(6) / 0.02	143(6) / 0.06	143(6) / 0.10	143(6) / 0.17	143(6) / 0.27	0.62
	3	(a)	Newton-SOR	147(8) / 0.01	150(11) / 0.05	152(13) / 0.07	153(14) / 0.13	154(15) / 0.19	0.45
			Newton-2EGSOR	83(8) / 0.00	84(9) / 0.02	85(10) / 0.04	86(11) / 0.08	86(11) / 0.12	0.26
			Newton-4EGSOR	58(7) / 0.00	59(8) / 0.02	59(8) / 0.07	60(9) / 0.07	60(9) / 0.11	0.24
		(b)	Newton-SOR	100(11) / 0.00	103(14) / 0.03	104 (15) / 0.06	105(16) / 0.11	106(17) / 0.17	0.37
			Newton-2EGSOR	54(7) / 0.00	55(8) / 0.02	56(9) / 0.04	57(10) / 0.07	57(10) / 0.10	0.23
			Newton-4EGSOR	37(6) / 0.00	39(7) / 0.01	39(8) / 0.03	39(8) / 0.06	39(8) / 0.09	0.19
		(c)	Newton-SOR	157(9) / 0.01	160(12) / 0.03	162(14) / 0.07	163(15) / 0.14	164(16) / 0.21	0.46
			Newton-2EGSOR	82(8) / 0.00	83(9) / 0.02	84(9) / 0.04	84(10) / 0.08	85(10) / 0.12	0.26
			Newton-4EGSOR	56(8) / 0.00	57(9) / 0.02	57(9) / 0.04	58(10) / 0.07	58(10) / 0.11	0.24

Detecting Depression in Videos using Uniformed Local Binary Pattern on Facial Features

Bryan G. Dadiz[1,2*] and Conrado R. Ruiz, Jr.[1]
[1] College of Computer Studies De La Salle University, Taft Manila, Philippines
[2] Technological Institute of the Philippines, Quiapo Manila, Philippines
bryan_dadiz@dlsu.edu.ph

Abstract. The paper presents the classification model of detecting depression based on local binary pattern (LBP) texture features. The study used the video recording from the SEMAINE database. The face image is cropped from a video and extracting the Uniformed LBP features in every single frame. Video keyframe extraction technique was applied to improve frame sampling to a video. Using the SVM with RBF kernel on the original ULBP features, result showed an accuracy of 98% on identifying a depressed person from a video. Also, part of the classification is to implement Principal Component Analysis on the original ULBP features to analyze facial signals by comparing both of the accuracy results. Using the original ULBP features with SVM applying radial basis function kernel, it resulted higher in accuracy whereas the result of using only ten features computed from the PCA of the original ULBP features. The result of the PCA decreased by 5% gaining only 93% in accuracy applying the same cost and gamma values of SVM RBF kernel used on the original ULBP features.

Keywords: Computer Vision, Local Binary Pattern, Facial Features, Depression Analysis

1 Introduction

The medical or emotional disturbance could be a mental temperament issue, caused by a person's trouble in adapting to distressing life occasions, and displays diligent sentiments of pity, negativity, and stress. In 2002 the World Health Organization (WHO) recorded major depressive disorder as the fourth most crucial reason for being handicapped within locality or around the world and anticipated it might be the second driving reason by 2030. [1]. As of late, World Health Organization estimated that more than 800 million people die from self-death or suicide consistently every year and no less than 20 times more attempted suicides. Suicide is the aftereffect of a deliberate act with the expectation to finish one's life. In the Philippines, Department of Health's National Center for mental health detected that the suicide rate for a male is 2.5% for every 100,000 groups of individuals and 1.7% for women. In the 2,500 suicide cases recorded in 2012, over 2,000 were male, and around 500 were female. There can be additional cases underreported attributable to shame, or dread of people with unsafe thought to be judged. However, there's an occasional suicide rate as con-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_40

trasted with other ASEAN nations. A close-by review by Perlas, Tronco et al. noticed that 5.3 percent of these studies were experiencing grief. Another report showed 4.5 million Filipinos are burdened with depression. Worldwide, the rate of depression ranges from 2.6 to 29.5 percent. Moreover, a rising distress incline is seen among the Filipino elderly. A crowded area in Rizal, Philippines scored 6.6 percent rate of depression utilizing the Geriatric Depression Scale, which showed that melancholy is present even in sound groups, agreeing to Business Mirror [2]. Hamilton Rating Scale [3] is proven a high quality level analytic evaluation instrument for depression however on the sentiment of individual clinicians on the other hand, there is also another tool which is the Suicide Probability Scales [4]. These methods give out a score to determine the level of sadness or a probabilistic occurrence of suicide. However, this is dependent on the patients' needs and trustworthiness to impart their indications, temperaments or comprehension. [5] As of the time being there are no goals or specific clinical measures for depression. Affect state detection has been a dynamic field of research over ten years. Notwithstanding, a restricted consideration is given towards appropriateness of these methods for programmed dejection examination. As indicated by the speculation proposed by Ellgring [6], depression prompts a momentous drop in facial movement, while with the change of subjective facial action.

Thinking about Ellgring's speculation as a beginning stage, in one of the original work towards automatic dejection investigation, a study [7] broke down the facial reaction of the subjects recorded on a video clip. Local shape and texture features were figured from each fifth video sequence and the method is called Active Appearance Models (AAM) [8].

In the area of facial recognition, the method LBP is becoming popular. The LBP operator [9] is the most outstanding method for texture descriptors; Utilization of the LBP operator has been successful for facial recognition on the results of T. Ahonen, [10] images of the face can be seen as a structure of small forms such as flat areas, spots, lines, and edges which can be well described by LBP. Also, PCA is highly used for improving classification by getting the Eigenvalues of the features extracted. [11] Furthermore, this project targets to deal with an effective way of computing depression and developing robust and non-invasive methodologies for suspecting a person's emotional state with a disorder for a mental state such as depression.

2 Introduction

Understanding the face signals for emotion analysis has been prevalent in computer vision and affective computing groups. Two decades have passed, various texture, geometric, static and temporal visual descriptors have been presented for various related expression analysis problems. Facial expression analysis methods can be broadly divided into three categories based on the type of feature descriptor used. One is shaped feature-based techniques utilizing geometric localization of face. Second, the class consists of appearance feature-based techniques, which analyze skin texture. The third is the hybrid method that uses both appearance and shape features [12-14]. Moreover, some applications of automated depression diagnosis were subtopics in the facial signal analysis. Critical areas in determining or diagnosis of depression are identifying behavioral indicators of distress, assessing variation of movement over time or slow of action and the effect of an intervention to people and their response [15]. The given approaches from previous studies all use high-dimensional audio-visual features, such as eye stabilization or gaze, movement of the body and head, speech, vocal pauses and quality of voice.

A study by Joshi et al. [14] proposed the analysis of intra-facial muscle action and the movements of the shoulders and head in a captured motion picture to analyze depression. The study computes Space-Time Interest Points(STIP) [16] and appearance features using Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [17]. Furthermore, for the audio analysis, frequency, loudness, intensity and mid-frequency cepstral coefficients were used (MFCC) was implemented. Using the aforementioned techniques with the classification algorithm SVM [18] the accuracy was up to 91.7% with the binary classification task (depressed vs. non-depressed).

On the other hand, many works have been done on analyzing facial expression and recognition, the diagnosis or automatic analysis of facial expression for depression is still an under-researched area. Another study of multimodal approach for assessing depression was the study of Dibeklioglu et al. [19] it computes for face, posture and vocal behaviors using logistic regression classifiers and leave-one-out cross-validation. The study infers that head movement and face outperforms vocal prosody however combination of the three is possible for detecting depression.

The demonstration of wearable sensors was introduced on the study of Fedora et al. [20] the goal is to measure the electro dermal activity (EDA) on the left and right palms of the subject that has a major depressive disorder while being subjected to Transcranial Magnetic Stimulations (TMS). The study shows that when validated by the Hamilton Depression Rating Scale (HRDS), the results followed the pattern of depression scores for EDA. The EDA on the right hand became more dominant or which is the dominant hand when the depression worsened. It inferred that the possibility to prevent depression by observing changes in an individual. The study showed promising results on early detection of depression. However, the reliance on wearable sensors for the hand can be of a bit problem. Ideally, a more natural and faster which is ubiquitous is ideal.

In the Audio Visual Emotion Challenge (AVEC) 2013 a platform is provided for aspiring researchers to participate in providing solutions or new approaches for utilizing audio, video/or physiological analysis of emotion and depression. The affective computing and social signal processing researchers are participating in this ongoing work on depression severity estimation coined as “behavioral edicts” and desire to help mental health practitioners by quantifying facial and vocal expressions. On AVEC 2013 there is a dataset provided for depression analysis called SEMAINE database by Queen's University Belfast. [21]

The study of Hyett et al. [22] conducted a study to understand the critical differentiation in connectivity patterns of the brain into plotting independent component analysis (ICA) maps on functional magnetic resonance imaging (fMRI). The study was conducted to melancholic patients to understand the causes and as a diagnostic tool. As a diagnostic tool fMRI equipment is limited, invasive and high at cost. Facial expression analysis has the potential to solve this flaws; the thing is that it more and more publicly available datasets are out on the internet for video analysis.

As of the studies mentioned above, while much more work has been done using multimodal analysis of depression, this study focuses on the binary classification of depression just like the study of Joshi et al. however this study focuses only on analyzing facial features on a video thus, audio is not a concern. In Addition, the study will analyze facial feature per each frame in a video and unlike the study of Shalini Bhatia et al. [23] instead of using LBP-TOP as the state of the art way of analyzing video using this study implements keyframe extraction technique called as thresh holding method inspired on the research of Sheena C.V. et al. [24] For each extracted keyframe ULBP will be extracted and used as features for classification. Furthermore, for the classification, the study will use the Support Vector Machine classifier with RBF kernel to assess if the method will be effective since the previous studies used SVM there is a better chance to compare the results of this study to previous works above. The focus of this research is (a) to investigate the pattern of depression computed on a face in a video clip implementing a robust texture analysis algorithm such as LBP operator (b) to develop a model to classify depressed individual using ULBP's Feature Vector and conclude an acceptable accuracy score. (c) Finding the optimal features of LBP using PCA analysis

3 Methodology

3.1 SEMAINE Database Description

The corpus used in the study for the purpose of detecting depression on facial analysis is the SEMAINE corpus [25]. The database is the recorded conversation between people and a virtual person. The original study of the corpus is to understand natural social signals that happen in exchanges of conversation with an artificial human or robot. The database is obtainable for research, analysis and case study from HTTP: //Semaine-db.eu. It includes someone interacting with showing emotion conventional characters. The state of affairs utilized in the documentation is named the Sensitive Artificial hearer (SAL) technique. The characters were Prudence, is neutral-tempered and sensible; Poppy, a contented and cheerful character; Spike, a character with a lot of resistance and has an angry tone; and Obadiah, is sad and depressive.

3.2 The Keyframe Extraction Method

The method of obtaining the keyframes to sample frames for LBP feature selection is computing the distinction between two consecutive frames. If the results of difference are bigger than the threshold, then take into account it as a keyframe. To explain further the initial step is to select a video then each frame is computed, and in every iteration, throughout the video, the output frame is extracted. The computation is that two frames are extracted and converted to greyscale, then their histogram difference is computed. The summation of the histogram elements is then calculated and to be returned. The variance and mean will be computed as the threshold. Also, for every iteration, this threshold value is computed and brought into comparison with the total value taken out for the histogram difference that was calculated previously. Whenever the Difference histogram of two images is greater than the value that is the threshold, then the following image is selected and becomes the keyframe. Finally, on the execution for every iteration, there will be a collection of key frames obtained which will be used to extract or segment the face. [36]

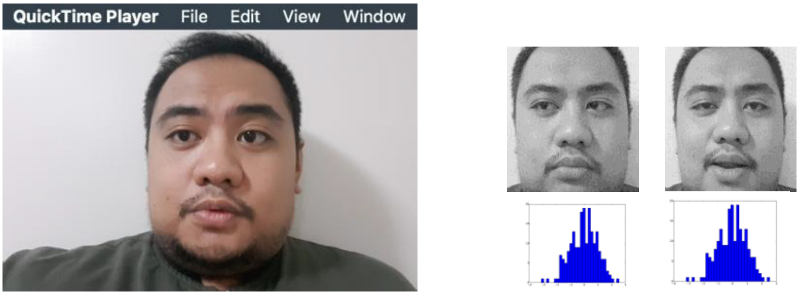


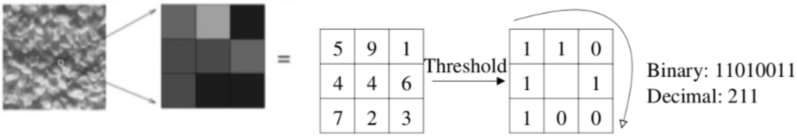
Fig. 1. The Difference of Each Keyframe Histogram

3.3 The Process of Face Segmentation

To perform Local Binary Pattern transformation on the full face as this study's region of interest, a facial detection algorithm was performed using the Viola and Jones [26] method. The face detection procedure classifies images based on the value of simple features. Viola-Jones Method is implemented in the study. The first two features optimally selected by the AdaBoost method. These features overlay the training face The horizontal box feature measures the distance of similarity in the intensity of the eye regions across the cheeks. It exploits on the observation that the eye section of the face is often darker than the cheeks. The much smaller but slightly vertical pattern features to compare the intensity of the nose bridge in the eye region.

3.4 Local Binary Pattern

The LBP texture description operator was intentionally for texture classification. The operation assigns a label to every pixel of an image by using a threshold of 3x3-neighborhood of each pixel with the center pixel value and considering the result as a binary number. Then, the pixel patterns are converted to a histogram that will be used as a texture descriptor. In previous researches, as seen on Fig. 2 the LBP operator was extended to manage textures at completely different scales in later extended the neighborhood in increasing sizes.



5	9	1
4	4	6
7	2	3

1	1	0
1		1
1	0	0

Binary: 11010011
Decimal: 211

Fig. 2. The Basic LBP Operator

Defining the native neighborhood as a collection of sampling points equally spaced on a circle focused on the element to be labeled permits any radius and range of sampling points. It needs to define the native neighborhood as a collection of sampling points that are equally spaced circulating the element to be labeled by a pattern, and this permits any radius and range of sampling points. Once the sampling purpose does not fall within the center, the usage of bilinear interpolation is required.

3.5 Classification

An empirical analysis of classifying depression from extracted facial expressions out of a video using the SEMAINE corpus is examined in this study. The results of the five machine learning techniques such as K Nearest Neighbors, Logistic Regression, Decision Tree, SVM and Multi Layered Perceptron will be analyzed to determine the best model for facial classification of depression. The overall knowledge discovery process, it is interactive and iterative involving, more or less, the following steps.

First step is to prepare the SEMAINE database for analyzing videos of depressed and not depressed classification based on full facial expression. The second step is to process the six video instances with the four emotional characters portrayed by the actors. In each video, keyframes are extracted using an extraction method. Third, determining features by getting the pixel-based feature value using the Uniformed Local Binary Patterns method. In Each key frame instances, ULBP features were extracted and saved as instances for training/validation. Then, PCA is used for dimensionality reduction of the 59 features extracted from each keyframe to gain the results using only the important features and its effect on the accuracy. Fourth, on validating the accuracy, the ten cross-fold validation for the training/testing dataset is used. Fifth, the supervised classification using the baseline algorithms for machine learning, e.g., K Nearest Neighbor, Decision Tree, Support Vector Machine, Logistic Regression and Multi-Layered Perceptron are used for training and results were analyzed. Finally, the accuracy of the model will be evaluated when dimension was reduced and recommendation of the experiment results.

4 Results and Discussion

The paper experiments and discussion shall be discussed in this chapter to show the scientific findings and patterns and inferences towards completing the experiments.

4.1 The LBP transformation of Sample Frames from SEMAINE Corpus

As seen in the images of Fig. 3 the LBP operator transformed the grayscale images showing evident facial features whereas it emphasizes the edges or lines of the face and regions like the eyes, mouth, nose, jawline and possibly face-orientation. In the experiment, it is suspected that the LBP operation can determine the face orientation by having a pattern for a semi occluded human face example facing a different angle as seen on Fig. 3(d). It is observed a semi occluded human face because the hair of the patient can cover the forehead and important changes on that part can be hidden and not be detected. The ULBP features of each frame extracted from a labelled video are saved as an instance for training the classification model.

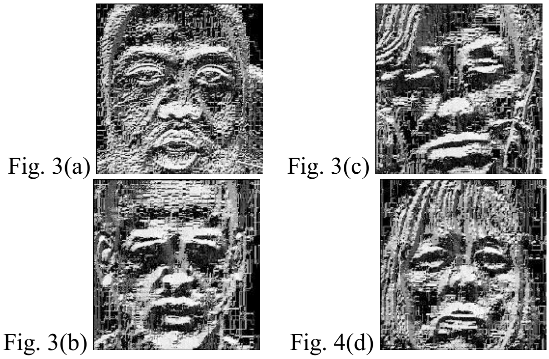


Fig. 3(a) Fig. 3(c)

Fig. 3(b) Fig. 4(d)

Fig. 3. Results of Transforming the Grayscale Keyframes with LBP operator

4.2 The LBP Histogram

The extracted keyframe's histogram is used as instances for training the machine. Eight examples videos from Obadiah (Depressed) was processed to provide ULBP histogram of each keyframe on the videos and labeled as Depressed. On the other hand, Poppy, Spike, and Prudence videos were processed the same as Obadiah but Labelled as Non-Depressed.

4.3 PCA Results

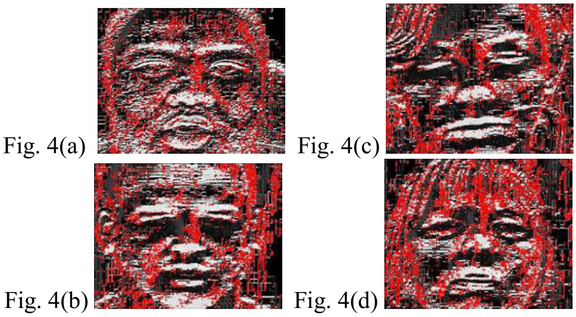


Fig. 4(a) Fig. 4(c)

Fig. 4(b) Fig. 4(d)

Fig. 4. Shows the images transformed by the LBP operator with inserted markers of Eigen features from PCA

PCA resulted in the Uniformed Binary Patterns features 60, 243, 120, 124, 126, 127, 128, 131, 135, 159 that were plotted to the LBP transformed frames to analyze where are the pixels that determine the classification of depression on full facial analysis within the video. Based on Fig. 4(a), a manifestation that there is a concentration on the nose, eyes, cheeks, mouth and some on the temple parts of the forehead. It also shows that a part of a face was selected provided by the PCA LBP features, looking closely on the image there is only a partial selection of the face at most sections, only half of the face shows a concentration of pixels. It can be observed that the hair participates in determining depression. However, this can be considered noise since hair movement cannot be exclusively caused by human movement alone. Moreover, there are also markers that can be seen not located on the face and the reason is that the segmentation has captured some parts of the background that may move or change color overtime due lighting, another suspected noise that affects the accuracy of the classifier. In contrast with Fig. 4(a) it was observed that the actual video shows that the subject often moves his head. Therefore, more pixels are concentrated on the face than outside of the face.

Table 1. Classification result using extracted face frames using ULBP

ULBP	**Classifier Performance**		
	Classifier	***Accuracy***	***F-Score***
	Decision Tree	69%	75%
	Logistic Regression	83%	83%
	k-Nearest Neighbor	98%	97%
	Neural Network	98%	97%
	SVM	68%	67%
	SVM-RBF	98%	97%

The experiment is conducted using several machine learning algorithms as shown in Table 1. Results have shown that the classifier SVM with RBF kernel has to be the highest accuracy that reached 98.58% in accuracy. It is also notable that Neural Networks could also be a potential classifier however training time for it requires a couple of hours. Using SVM classifier, the result of each keyframe histogram features for each session showed a high accuracy result where the RBF kernel values for cost is 1.0 and gamma is 100. Although when observing the results of Table II. There is a decrease of accuracy result using the same model going to 5% lower from original ULBP features, nevertheless still a good result of accuracy whilst using only ten ULBP features.

Table 2. Classification result using extracted face frames using ULBP PCA Eigen features

ULBP-PCA	**Classifier Performance**		
	Classifier	***Accuracy***	***F-Score***
	Decision Tree	70%	75%
	Logistic Regression	70%	69%
	k-Nearest Neighbor	91%	90%
	Neural Network	92%	92%
	SVM	64%	66%
	SVM-RBF	93%	92%

5 Conclusion and Recommendation

In summary, the study shows the pattern of detecting depression in a video using ULBP and PCA. (a) It shows that full face analysis of depression in a video is focused on the part of eyes, nose, mouth, cheeks and head temple. (b) The model for depression classification on full facial features accuracy resulting in 68% using SVM classifier with the normal linear kernel and whereas SVM using Radial Basis Function as a kernel yielded 98% a significant increase in accuracy. (c) On the other hand, PCA resulted features from ULBP descriptor has shown a 5% decrease on its performance lower than the original, using only ten features. The usage of the original ULBP features provided high result of accuracy than the ten features that the PCA selected. There is a loss on the model's performance using the same RBF kernel values. Nevertheless, the accuracy is still highly acceptable for determining depression using ULBP facial features on a video. On the other hand, when segmenting the face from video frames using Viola and Jones method, it resulted to capture some parts of the background along with the face. Thus there is a recommendation to have a more robust segmentation method to capture absolute part of the face excluding further noise from the background.

References

1. Mathers, C.D., Loncar, D., (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 3, 2011–2030.
2. Chrisha Ane Magtubo. (2017). MIMSToday: World Health Day 2017 focuses on depression and suicide. Retrieved from https://today.mims.com/world-health-day-2017-focuses-on-depression--suicide.
3. Hamilton, M. (1960). The Hamilton Depression Scale—accelerator or break on antidepressant drug discovery. *Psychiatry*, *23*, 56-62.
4. Cull, J.G., Gill, W.S., (1982). Suicide probability scale. In: Western Psychological Services. Western Psychological Services, Los Angeles, CA, pp. 1997–2005.
5. Mundt, J.C., et al. (2012). Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psych. 72, 580–587.
6. H. Ellgring, (2008). Nonverbal communication in depression. Cambridge University Press,
7. G. McIntyre, et al. (2009). "An Approach for Automatically Measuring Facial Activity in Depressed Subjects," ser. ACII'09,
8. J. Saragih and R. Goecke, "Learning AAM fitting through simulation," *Pattern Recognition*, vol. 42, no. 11, pp. 2628–2636, 2009.
9. T. Ojala, M. Pietika¨inen, and D. Harwood, (1996)."A Comparative Study of Texture Measures with Classification Based on Feature Distributions," Pattern Recognition, vol. 29, no. 1, pp. 51-59,
10. Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence, 28(12), 2037-2041.
11. A. Pentland, B. Moghaddam, and T. Starner, (1994). "View-Based and Modular Eigenspaces for Face Recognition," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, pp. 84-91,

12. M. Hayat and M. Bennamoun.(2014). An automatic framework for textured 3d video-based facial expression recognition. *IEEE Transactions on Affective Computing*, 5(3):301–313,
13. M. Hayat, M. Bennamoun, and A. El-Sallam. (2012). Evaluation of spatio-temporal detectors and descriptors for facial expression recognition. In *5th International Conference on Human System Interactions (HSI)*, pages 43–47,
14. Joshi, J., Goecke, et al. (2013). Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interf. 7, 217–228.
15. J.M. Girard and J.F. Cohn. (2014). Automated audiovisual depression analysis. *Current opinion in psychology*, 4:75–79,
16. I. Laptev. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123,
17. G. Zhao and M. Pietikainen. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915– 928,
18. C. Chang and C. Lin. Libsvm: (2011). A library for support vector machines. software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27:1– 27:27,
19. H. Dibeklioglu, Z. Hammal, Y. Yang, and J.F. Cohn. (2015). Multimodal detection of depression in clinical interviews. In *ACM International Conference on Multimodal Interaction*, pages 307–310,
20. S. Fedor, P. Chau, N. Bruno, R. Picard, and J. Camprodon. (2016). Asymmetry of electrodermal activity on the right and left palm as indicator of depression for people treated with transcranial magnetic stimulation. In *Annual Meeting of the Society of Biological Psychiatry (SOBP'16), Atlanta, Georgia*,
21. Valstar, M., Schuller, B., et al. (2013, October). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* (pp. 3-10). ACM.
22. M. Hyett, M. Breakspear, K. Friston, C. Guo, and G. Parker. (2015). Disrupted effective connectivity of cortical systems supporting attention and inte- roception in melancholia. *JAMA psychiatry*, 72(4):350–358,
23. Bhatia, S., Hayat, M., Breakspear, M., Parker, G., & Goecke, R. (2017, May). A video-based facial behaviour analysis approach to melancholia. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on* (pp. 754-761). IEEE.
24. Sheena, C. V., & Narayanan, N. K. (2015). Keyframe extraction by analysis of histograms of video frames using statistical methods. *Procedia Computer Science*, *70*, 36-40.
25. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. (2012). The Semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing, 3:5–17,
26. Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

Malicious Software Family Classification using Machine Learning Multi-class Classifiers

Cho Cho San, Mie Mie Su Thwin, Naing Linn Htun

Cyber Security Research Lab, University of Computer Studies, Yangon, Myanmar
chochosan@ucsy.edu.mm, drmiemiesuthwin@ucsy.edu.mm,
nainglinnhtun@ucsy.edu.mm

Abstract. Due to the rapid growth of targeted malware attacks, malware analysis and family classification are important for all types of users such as personal, enterprise, and government. Traditional signature-based malware detection and anti-virus systems fail to classify the new variants of unknown malware into their corresponding families. Therefore, we propose malware family classification system for 11 malicious families by extracting their prominent API features from the reports of enhanced and scalable version of cuckoo sandbox. Moreover, the proposed system contributes feature extraction algorithm, feature reduction and representation procedure for identifying and representing the extracted feature attributes. To classify the different types of malicious software Random Forest (RF), K-Nearest Neighbor (KNN), and Decision Table (DT) machine learning multi-class classifiers have been used in this system and RF and KNN classifiers provide 95.8% high accuracy in malware family classification.

Keywords: Malware analysis, Malware classification, Random forest, k-Nearest neighbor, Decision Table

1 Introduction

For years, a number of new malware have exponentially increased, which creates difficulty for malware analysts and anti-virus vendors to profile the families, as they need to extract information out of this large-scale data. Although many anti-virus vendor companies such as Microsoft, Avast, BitDefender and Kaspersky provide for detecting and profiling the malicious samples, the number of malware is growing more than ever in recent years together with their seriousness. Malware continues to be a blight on the threat landscape with more than 357 million new variants observed in 2016 [1]. The goal of malicious software analysis is to gain an understanding of the behaviors and functions of a malware so that the analysts can build the defense procedure to protect an organization's sensitive information and network postures. A number of new malware variants are very effective for evasion the anti-virus and anti-malware applications, and malware creators often use the similar method of attacking to build new malware in a short period of time. Therefore, distinguishing and classifying the nature of malicious executable from each other is very important because the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_41

polymorphic and metamorphic malware are increasing and attackers may use them as long as there is a profit to be made.

One of the main goals of the proposed system is to identify the malicious traces reliably, i.e., to decrease false positive detections, which is the major challenge of behavior-based anomaly detection. So, the proposed system performs the malicious family profiling on over 10500 samples from 11 different families using machine learning classifiers in WEKA[1]. This paper highlights the malicious features such as API sequences in terms of their arguments such as registry, function call arguments and system call with their relevant families. This paper also proposes a feature extraction algorithm along with the feature reduction and representation process for malicious family classification system. The approach provides the best classification performance and good accuracy with nearly 96% for training data and 11 test datasets yields good results over 90%.

The rest of the paper is organized as follows. Section 2 provides the literature review and section 3 highlights the proposed malware family classification system. Section 4 shows the experimental results and discussion respectively. Finally, the conclusion and the outline of the future research plans are described in section 5.

2 Literature Review

Malware analysis can generally be classified into static, dynamic analysis and hybrid analysis approach. The static feature uniquely identifies the signature of malware or malware families. Basic static analysis techniques such as scanning with anti-virus software, looking at the malware with a hex editor, unpacking the malware, performing a strings search and disassembling the malware. Static analysis is vulnerable to code obfuscation techniques [13]. In particular, polymorphic malware has a static mutation engine that encrypts and decrypts the code, while metamorphic malware automatically modifies the code each time it is propagated [7]. In dynamic analysis behavior of malicious software is monitored in emulated environment and traces are obtained from the reports generated by sandbox. Although it can deal with code evasion techniques, more dangerous than the static analysis [13]. Nowadays, malware is often equipped with various code obfuscation techniques, making pure static analysis extremely difficult for the anti-virus vendors, malware analysts and researchers. By actually executing the malware, dynamic analysis can overcome these code obfuscation techniques. This is because no matter what code obfuscation methods the malware is equipped with, as long as the malware exhibits the malicious behaviors during the dynamic analysis, malware analyzer can observe and analyze these malicious behaviors [2]. Although virtualization of the malware lab is great for cost reduction, there are issues with using virtualization software. Some of the more sophisticated malware today will attempt to detect a VM. If the malware detects it is being run on a VM, it will not execute [6]. We counted and noted that these kinds of malware that

[1] https://www.cs.waikato.ac.nz/ml/weka/

evade from execution and remaining silent in analysis environment for future research work by performing the hybrid approach.

In [3] used a subset of 126 APIs from 6 most important dynamically linked libraries (DLL). The authors also used feature selection techniques to lower the number of features. When the authors combined AdaBoostM1 with J48 classifier, it showed the best performance on the accuracy of 98.4% for their dataset. In [11], API behavioral traces were obtained using a modified version of Cuckoo sandbox and their approach relies on random forest machine learning classifier for performing both malware detection and family classification by extending their previous work [14]. In [4] and [5] classified malware from benign by extraction API calls with a small number of samples. In [12], 1,086 malwares from 7 malware classes were used in their approach and they conducted a five-fold cross-validation for the evaluation. As for classification they used Decision Tree classifier with the precision of 83.60%, and the SVM classifier with the precision of 88.30%. In [15], five malware classes for total 2000 samples were used for experiment, n-gram was used to extract API call sequence patterns by hooking the user level. They divided 5-class problems into five 2-class problems similar to our work such as Worm vs rest, Backdoor vs rest, Trojan-Dropper vs rest, Trojan-Downloader vs rest and Trojan-Spy vs rest. In this section we present the state-of-the-art approaches that use static and dynamic analysis to perform either malware detection or family classification. In the aforementioned research articles, API calls are the mainly used as parameter for creating features. The previous works mentioned above have some weaknesses in the small number of tested samples or family in the classification and detection system although some related works yield high accuracy. In our proposed system, we tested over 10000 samples from multiple malicious categories and extracted API features parameters such as system, network, process, file, and registry are chosen as features in the categories of API calls.

3 The Proposed Malware Family Classification System

One of the main reasons to contribute this research is to classify malware family as malware developers are using the polymorphic, metamorphic, packing, and encryption techniques, which cause the high volume of malware samples. In this proposed system, we consider the malware that have used packing or encrypting techniques. The longer the malware can remain undetected on a victim or compromised machine, the more the cybercriminal or attacker can profit. For the above reasons and problems, therefore, prominent features extraction from generated report files in malware analysis is very important for detection and prevention from malware executing and infecting. So, we propose the Malware Feature Extraction Algorithm (MFEA) to point out the dominant features based on the generated Java Script Object Notation (JSON) report files. The proposed system also contributes Feature Representation for the presence and absence of features in the extracted files along with the features reduction procedures performed on the feature space for the classification efficiency. The process flow diagram shows the proposed step by step procedure of the malware analysis for family classification architecture in Figure1 together with the following detailed steps.

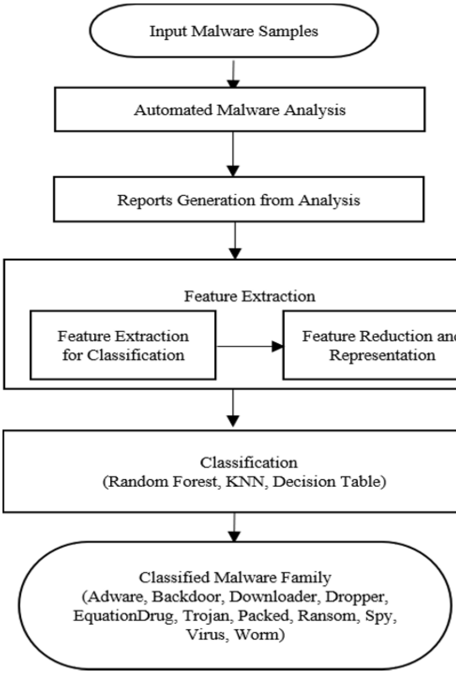

Input Malware Samples
Automated Malware Analysis
Reports Generation from Analysis
Feature Extraction
Feature Extraction for Classification
Feature Reduction and Representation
Classification (Random Forest, KNN, Decision Table)
Classified Malware Family (Adware, Backdoor, Downloader, Dropper, EquationDrug, Trojan, Packed, Ransom, Spy, Virus, Worm)

Fig. 1. Overall architecture of malware analysis and classification system.

3.1 Collecting Malicious Samples

Although we collected a large amount of sample from virus share[2], we tested over 10000 malicious samples in this work. Total 9068 samples from 11 malware families that experiment in our research work are described in Table 1. The class labels are also described together with the families and the number of sample per family.

Table 1. Malware family and the total number of samples per family

Class Label	Malware Family	Total number of samples	Class Label	Malware Family	Total number of samples
1	Adware	2354	7	Ransom	490
2	Backdoor	764	8	Virus	928
3	Downloader	1010	9	Spy	582
4	Dropper	352	10	Trojan	589
5	EquationDrug	158	11	Worm	867
6	Packed	974			

[2] http://tracker.virusshare.com:6969/

3.2 Automated Malware Analysis

The proposed system performs dynamic analysis in the secure virtual environments with Window-7 Operating System (OS) as guest OS for analyzing the malware sample in Virtual Box[3] and Ubuntu as host OS. In this phase, cuckoo[4] sandbox will be used as automated malware analysis system in the proposed framework. It is widely used by academic and independent researchers as well as small to large companies and enterprises. In the analysis phase, some malware evades from the analysis because of obfuscation or polymorphic nature. The proposed system notes that this kind of malware and later we will perform for further extension.

3.3 Generating Reports from Analysis

This phase describes the reporting of the output from analysis. It generates the analysis result with HTML, JSON format as report. However, in our proposed system, we use JSON report format for extracting malicious features.

3.4 Feature Extraction

To extract the labels and prominent malware features, we propose the Malicious Feature Extraction Algorithm (MFEA) to extract API calls from malware samples. The features are based on API calls and their input arguments such as process, registry, system, file, and network features. In this system we only use API features with their arguments. Total 1732027 API features have been extracted in our proposed system for 10500 malware samples. The proposed feature extraction algorithm describes in Algorithm 1. As for feature extraction, we define R as a set of JSON report files and F contains a set extracted API features files for each malware sample. And APIDB defines as a database of total global API features used by malware samples.

```
Algorithm 1: Malicious Feature Extraction Algorithm (MFEA)
Input: R contains a collection of JSON report files $R_i$
Output: F contains API features file $F_i$, APIDB
API = {file, registry, process, system, network}
while R do
  for each API in R do
     if API exists in R then
              Extract this API from R;
              F += API;
       if this API is not existing in APIDB
         APIDB += API
       end if
   end if
  end for
```

[3] https://www.virtualbox.org.

[4] https://cuckoosandbox.org

```
11: end while
```

3.5 Feature Reduction and Representation

The features are extracted according to the sequence of calls by malware without using n-gram technique and sorting. The total number of extracted features are 1732027 API, which is quite large for classification. In this case, feature reduction process is very important to reduce the number of feature without losing the classification performance. API features that meet the following four conditions, which can affect the classification system in terms of accuracy and processing time. Therefore, the proposed system also contributes the feature reduction and representation on extracted features to improve the classification performance. Feature reduction steps will perform if it meets the following conditions:

1. Extracted feature files contain duplicate features
2. Features contain special characters (e.g. double code, dollar sign)
3. Features start and end with noise data (e.g. underscore or hash sign, etc.)
4. Features with wrong spelling and duplicate character

After this stage, 1417 API attributes remain from the total extracted API features. We perform the classification on 11 families and discard some malware family that do not have at least 150 samples. So, the total tested number of samples are 9068 with 11 families. After performing the feature reduction step, the output data will represent into their relevant feature representation. Since the extracted feature set, containing the failed and successful APIs, is quite large, we have to find a way to describe it in a clear, compact and simple representative way. And for an API feature vector representation to ensure each API is existing or not, binary feature vector space is created as follows:

$$\mathrm{API_i} = \begin{cases} 1, & \mathit{if\ each\ API\ is\ in\ File} \\ 0, & \mathit{otherwise} \end{cases}$$

If the API features from each instance or extracted API feature file Fi contains in APIDB, this API feature will be defined as 1 and otherwise 0 will be defined in this API for each instance. For example,

Sample S1 = {1,1,1,0,1,0,1,0,0,1, ….}

After performing the feature reduction and representation processes, the resulted 1417 API attributes are stored in CSV format. And then we convert the CSV format into ARFF for classifying the dataset using Weka. We prepared the dataset after performing the feature extracting, and reduction by representing malware sample class label as 1 to 11 respectively with 9068 rows for malware samples in terms of API calls.

3.6 Family Classification using Machine Learning Multi-class Classifiers

After that machine learning algorithms are applied on our extracted API features to validate the effectiveness of the extracted features. Machine learning is largely divided into supervised learning and unsupervised learning. Supervised learning uses the correct input-output pairs as training data. The purpose of a supervised learning is to obtain a correct output against the input data. On the other hand, the purpose of an unsupervised learning is to find the regularity from input data [9]. This research area has a classification problem that the answer exists, so we use supervised learning algorithms. For malware family classification, the three machine learning algorithms such as Random Forest (RF), k-Nearest Neighbor (k-NN), and Decision Table that can classify the multi-class in WEKA to be used in this study.

1. Random Forest. Random Forest is an ensemble learning algorithm which multiple decision trees as a weak classifier combined to a classifier [9]. The Random Forest is appropriate for high dimensional data modeling because it can handle missing values and can handle continuous, categorical and binary data. The bootstrapping and ensemble scheme makes Random Forest strong enough to overcome the problems of over fitting and hence there is no need to prune the trees. Besides high prediction accuracy, Random Forest is efficient, interpretable and non-parametric for various types of datasets [10].

2. K-Nearest Neighbors (NN). Nearest Neighbors is one of the simplest, and also known as K- Nearest Neighbors (k-NN) algorithm. This algorithm assumes all instances correspond to points in the n-dimensional space. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. In nearest-neighbor learning the target function may be either discrete-valued or real-valued. The k-NN algorithm is easily adapted to approximating continuous-valued target functions [16]. In real world problems, data rarely obeys the general theoretical assumptions, making non-parametric algorithms a good solution for such problems.

3. Decision Table. A decision table with a default rule mapping to the majority class. This representation, called DTM (Decision Table Majority), has two components: a schema which is a set of features that are included in the table, and a body consisting of labelled instances from the space defined by the features in the schema. Given an unlabeled instance, a decision table classifier searches for exact matches in the decision table using only the features in the schema (note that there may be many matching instances in the table). If no instances are found, the majority class of the DTM is returned; otherwise, the majority class of all matching instances is returned [8].

The purpose of this work was to determine the best feature extraction and reduction, feature representation, and classification methods that result in the best accuracy. To profile the malware class label such as Trojan, or Adware, etc., we use the VirusTotal[5] by defining the majority vote of the sample's name because only one anti-virus vendor does not satisfy to label the sample. Sometime one anti-virus vendor can detect and profile some malware but it might not detect the others. So, we label the sample by using the highest majority vote from Virus Total result.

[5] https://www.virustotal.com

4 Results and Discussion

In this proposed system, over 10500 malicious samples with 11 different families are experimented. However, some malware evades from executing in the virtual environment and some family do not have enough samples at least 150, that's why total malware samples are 9068 samples. After analyzing the malicious samples, we extract the prominent features from the analysis' report using proposed feature extraction algorithm. After extracting the dominant API features, we perform feature reduction and representation for malware classification by implementing the Python code.

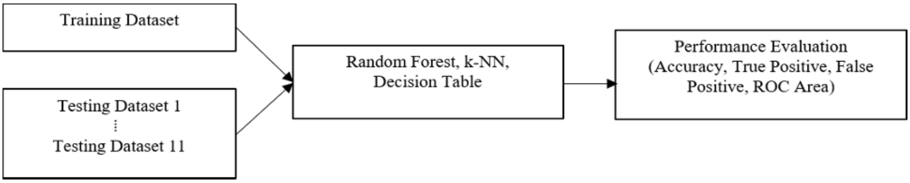

Training Dataset
Testing Dataset 1
⋮
Testing Dataset 11
Random Forest, k-NN, Decision Table
Performance Evaluation (Accuracy, True Positive, False Positive, ROC Area)

Fig. 2. Malware family classification on training and testing datasets

Fig. 2 shows the classification process for training and testing dataset. Most of the malware characterization and family classification works have conducted by splitting the training set into 10-fold cross validation or one testing dataset contains multiple families but not on the individual family testing dataset. In our system, we use this kind of approach in testing phase, OnevsAll (OvA), to validate our extracted API feature can correctly classify on their families or not. In our classification, we use machine learning algorithms such as random forest, k-nearest neighbor, and decision table algorithms to classify multiple malicious families.

Table 2. Family classification experiment on training dataset with and without cross validation.

	Without Cross validation				With Cross validation (cv = 10)			
Classifiers	Accuracy	TP	FP	ROC	Accuracy	TP	FP	ROC
Random Forest	0.958	0.958	0.005	0.999	0.846	0.846	0.016	0.979
k-NN	0.958	0.958	0.005	0.999	0.836	0.836	0.017	0.961
Decision Table	0.810	0.810	0.033	0.973	0.751	0.751	0.050	0.956

Random forest, k-NN produce same accuracy on training set with 95.8% and decision table provides 81% accuracy in Table 2. True Positive (TP) rate and False Positive (FP) rate are used to validate our extracted prominent malware features API dataset. With cross-validation (10) testing, random forest produces better accuracy than the other two classifiers. In testing phase, random forest classifier provides slightly better than the other two classifiers and it gives low FP rate. The malware family classification over 11 test datasets using random forest classifier for correctly and incorrectly classified instances evaluation performance in accuracy also describes in Table 3.

Table 3. Result of testing phase using random forest classifier.

Test Datasets	Accuracy Score	Test Datasets	Accuracy Score
Adware	99.80%	Ransom	94.30%
Backdoor	89.30%	Virus	97.30%
Downloader	93.40%	Spy	95.40%
Dropper	87.50%	Trojan	90.30%
EquationDrug	96.80%	Worm	97.70%
Packed	97.40%		

In testing phase, we divide 11 testing datasets based on the malware families to validate the extracted prominent API features which can choose the correct malware family. Moreover, these classification results indicate that there exist class specific signatures for every class which can be extracted. It is found that RF classifier is slightly better than the k-NN in test datasets. Both classifiers yield better accuracy in training phase than Decision Table for multi-class classification systems and high dimensional dataset and is therefore more accurate than the other classifiers. Correctly prediction percentage has been used to measure the accuracy.

For accuracy comparison with the related work mentioned above, our approach gives nearly 96% classification accuracy for training dataset and most of the tested datasets yield the best performance although the two test datasets, backdoor and dropper, are under 90% in accuracy. Random forest classifier not only provides the accuracy nearly 96%, but also results minimum number of incorrectly classified instances than the k-NN classifier, so it is the best classifier for our test dataset because of the nature of multi-class classification. However, k-NN also provides high classification performance in some test dataset than random forest and decision table classifiers.

5 Conclusion and Future Work

With an increasing amount of malware adopting new variants technologies to evade the current antivirus software or anti-malware detection systems, further research into defenses against serious targeted attacks is essential. Therefore, this system proposed an applicable dynamic malware software family classification framework and feature extraction algorithm for Cyber Crime Investigation System. Moreover, the feature reduction is one of the essential parts of the classification for performance and accuracy in malware classification. So, this system highlights their relevant and prominent features reduction system by using the proposed Malware Feature Extraction Algorithm (MFEA) and classifies the malicious family using Machine Learning Techniques such as random forest, k-nearest neighbor, and decision table classifiers in weka. The proposed system contributes the Feature Reduction and Feature Representation along with identifying and classifying between different malware family attributes. Beyond that, we perform the malicious family classification. The approach provides good accuracy for training dataset with nearly 96% and also provides best accuracy for the test datasets. The extracted API features dataset by using our proposed feature extraction algorithm covers the malicious family classification system to classify multiple families. For future work, we will take into account other features such as Dynamic Link Library (DLL) and static features in our future work. Moreover, we will also add more malware such as ransomware, spyware, APTs and clean ware samples by analyzing and extracting features for future malware family classification and detection system by extending the current work. Malware Detection system will also perform in our further extension by optimizing the malicious and benign features from dynamic analysis.

References

1. Internet Security Threat Report, Volume 22, Symantec (April 2017)
2. Yin, H., Song, D.: Automatic Malware Analysis: An Emulator Based Approach, SpringerBriefs in Computer Science, DOI 10.1007/978-1-4614-5523-3 7 (2013)
3. Salehi, Z., Ghiasi, M., Sami, A.: A miner for malware detection based on API function calls and their arguments, In: Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium on, pp. 563–568 (May 2012)
4. Uppal, D., Sinha, R., Mehra, V., Jain, V.: Malware detection and classification based on extraction of api sequences, In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2337–2342 (September 2014)
5. R. Tian, R. Islam, L. Batten, and Versteeg, S.: Differentiating malware from cleanware using behavioural analysis, Malicious and Unwanted Software (MALWARE), 5th International Conference on, vol. 5, no. 5, pp. 23–30 (2010)
6. Dennis Distler, Malware Analysis: An Introduction, SANS Institute, (December 14, 2007)
7. Ahmadi, Mansour, Dmitry, U., Stanislav, S., Mikhail, T., Giorgio, G.: Novel feature extraction, selection and fusion for effective malware family classification. In: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 183-194. ACM (2016)
8. Kohavi, R.: The power of decision tables. Machine learning: ECML-95, 174-189, (1995).
9. Kawaguchi, N., Omote, K.: Malware function classification using APIs in initial behavior. In: Information Security (AsiaJCIS), 10th Asia Joint Conference on, pp. 138-144. IEEE, (2015)
10. Qi, Y.: Random Forest for bioinformatics, http://www.cs.cmu.edu/
11. Hansen, Steven, S., Thor Mark Tampus, L., Matija, S., Jens Myrup, P.: An approach for detection and family classification of malware based on behavioral analysis. In Computing, Networking and Communications (ICNC), International Conference on, pp. 1-5. IEEE, (2016)
12. Hong, J., Park, S., Kim, SW.: On exploiting static and dynamic features in malware classification. In: International Conference on Big Data Technologies and Applications (pp. 122-129). Springer, Cham (Nov 17 2016)
13. Ranveer, S., Hiray, S.: Comparative analysis of feature extraction methods of malware detection, International Journal of Computer Applications. 120(5) (Jan 1 2015)
14. Pirscoveanu, Radu, S., Steven Hansen, S., Thor MT, L., Matija, S., Jens Myrup, P., Alexandre, C.: Analysis of malware behavior: Type classification using machine learning. In Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), International Conference on, pp. 1-7. IEEE, (2015)

15. S. Gupta, H. Sharma, S. Kaur, Malware characterization using windows API calls sequences, In: International Conference on Security, Privacy, and Applied Cryptography Engineering, Springer, Cham, pp. 271-280, (2016 Dec 14)
16. TM. Mitchell, Machine learning. WCB. (1997).

Modification of AES Algorithm by Using Second Key and Modified SubBytes Operation for Text Encryption

Aye Aye Thinn and Mie Mie Su Thwin

Cyber Security Research Lab, University of Computer Studies, Yangon, Myanmar
ayeayethinn@gmail.com, drmiemiesuthwin@ucsy.edu.mm

Abstract. Ciphering algorithms play a main role in this digital era, especially when data are transferred via internet. Many algorithms has been developed and used to encrypt and decrypt data for secure storing or transmission. At present, both synchronous and asynchronous encryptions are used to achieve the high security and to speed up the encryption time and process. Advanced Encryption Standard (AES) plays a prominent role among the synchronous encryption algorithms. It is also known as Rijndael Algorithm. Because of high performance of AES algorithm, it has been chosen as a standard among symmetric cipher algorithms. In this paper, we would like to propose a symmetric encryption algorithm. Modification is based on AES and we add an additional or second key. Another modification is also done at SubBytes step by adding the transportation operation in the original SubBytes operation. To analyze the performance of the modified proposed algorithm, Java language is used to implement the algorithm and then the performance is analyzed. After analyzing and verifying the experimental results, the proposed revised algorithm also shows good performance and high security from the cryptographic point of view. Based on the results of comparison between modified AES and original AES algorithm, our proposed algorithm can be used as a symmetric encryption algorithm, especially for the applications that share sensitive data files via insecure network.

Keywords: Encryption, AES, Rijndael, Cryptography

1 Introduction

The world is moving towards digital world and every country is moving its own pace for digitization and digital transformation. Using digital data via internet or network is becoming part of everything we do in our daily life and now is changing the way we live and work. Global digital transformation is accelerating the transformation of business and government, and even for the individual lives. Together with the digital transformation, more and more information are transferred or shared through internet and intranet. As a result, information security becomes even more important and vital for all of us when sensitive, confidential and valuable information are delivered via insecure network or store as archived data.

There are some basic requirements to delivery the electronic data or documents securely and safely. If we look from view of the sender of the information, ensuring the

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_42

integrity and confidentiality of information is a desire requirement. As for the information receiver, the non-repudiation and integrity are important aspects to achieve.

Security of information systems could be implemented with many widely known security algorithms, which can be adjusted with different settings for these algorithms. There are a lot of factors for security settings, with main important factors like the type of cipher, which prove security functionality, the processor time consumption, the size of packets, the general power consumption, the data type used and the battery power consumption [1].

All the symmetric and asymmetric encryption techniques are used to achieve the confidentiality of the information. A single key is used in symmetric encryption and all individuals who will receive the message must possess that secret key in order to communicate safely. Asymmetrical encryption uses a pair of keys, a private key and a public key. By using the digital signature, asymmetrical encryption technique can ensure the integrity and non-repudiation. The Advanced Encryption Standard (AES), also known as Rijndael, at the moment has not been broken. However, the cryptanalysis of AES has not stopped and many researches are finding new approaches to allow us to obtain competitive performance. Literatures [1- 3], [5] and [8-11] describe design and implementation of AES for improvements.

This research intended to propose a solution to encrypt and decrypt text files when they are transferred via insecure network. Encryption algorithm is based on AES algorithm. Modifications are made by adding an additional or second key to the AES and another modification at the SubBytes operation. This paper is organized as follows. The brief introduction is described in Section 1. Related work, Advanced Encryption Standard (AES) and Modified AES Algorithm (AES-R) are presented in Sections 2, 3 and 4 respectively. The Experimental Result is presented in Section 5 and conclusion of the paper is described in Section 6 together with proposed further research.

2 Related Work

Many published papers write about improving the performance of AES.

Ibtihal Mohamed A. Fadul and Tariq Mohamed H. Ahmed [2] proposed new ways to enhance the security of AES by using two secret keys. The additional key is used in both encryption and decryption to increase the security strength. The proposed method also retains the performance close to original AES algorithm as much as possible. Chittaranjan Pradhan and Ajay Kumar Bisoi [3] have proposed a solution by focusing on the security of the key used. For the key encryption, they used the 1-D Logistic Chaotic equation, cross-chaotic equation and the combined version of these two equations. Reena Mehla and Harleen Kaur [4] suggested work is focused on the amendment in the key expansion and shift row transformation of AES in order to make the respective algorithm more robust towards attacks. Their proposed work also reduced the encryption time taker for images and gives the better performance than AES. Their work can also contribute for more bandwidth efficiency. For adaption of image cryptography, Abdulkarim A. Shtewi, Hasan and Hegazy [5] offered an efficient modified AES. Their modified AES proposes shift row adjustment in AES, in order to give better results in terms of encryption time and security. Kazys Kazlauskas, Robertas Smaliukas and Gytis Vaicekauskas [6] examined and modified the AES algorithm to reduce the calculation of algorithm and to get improvement in data transmission. Their proposed method used parity bit generator to reflect a high level security and to achieve better data transmission without using Mixcolumns. Sumira Hameed, Faisal Riaz, Moghal, Akhtar, Ahmed and Ghafoor Dar [7] also worked to modify AES by taking the advantage of DES algorithm. They modified the AES by using the Permutation step rather than MixColumns step. The proposed algorithm is designed both for text and images encryption.

3 Advanced Encryption Standard (AES)

Advanced Encryption Standard (AES) is a symmetric-key block cipher algorithm and it is the standard of United States Government's Federal Information Processing. The design and strength of all key lengths of AES algorithm were sufficient to protect classified information up to SECRET level and TOP SECRET information (either the 192 or 256 key length). The key length of AES may be of 128, 192 or 256 bits but it uses a fixed block size of 128 bits. The input bytes array is first mapped to the State array at the beginning of encryption or decryption. Finally, the final value of State is mapped to the output array bytes [4]. The number of rounds in AES depends on the length of the key. There are 10 rounds for 128-bit keys and 12 rounds for 192-bit keys in AES algorithm. 256-bit keys will perform 14 rounds. Different 128-bit round keys are calculated from the original AES key and used each of these rounds. Each block in AES performs four primary processes, namely SubBytes, ShiftRows, MixColumns and AddRoundKey.

Decryption of AES cipher text is reverse operations of encryption process and subkeys are also in reverse order. During the decryption, each round consists of the four processes conducted in the reverse order − InvAddRoundKey, InvMixColumns, InvShiftRows and InvSubBytes. Since the decryption is the reverse of the encryption process, decryption algorithm is separately implemented, although they are very closely related [12]. AES has been widely adopted in today cryptography because of its support in both hardware and software. At present, the AES has not been broken and it can be only broken by the brute force or exhaustive search. The AES algorithm was designed to make it difficult to break by linear and differential analysis [10]. AES makes difficult for exhaustive key searches due to having flexible key length.

4 A Modified AES Algorithm (AES-R)

The overall encryption and description process of our proposed revised AES algorithm (AES-R) is shown in Fig. 1 and Fig. 2.

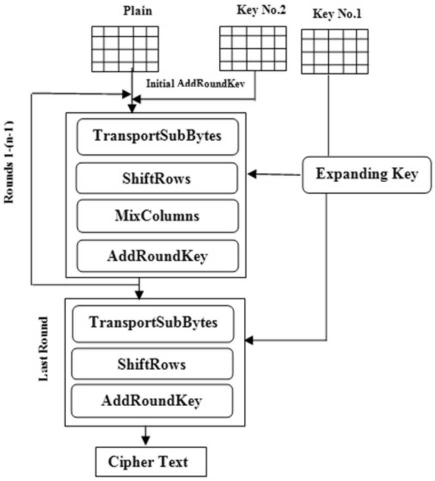


Fig. 1. Encryption Process of AES-R

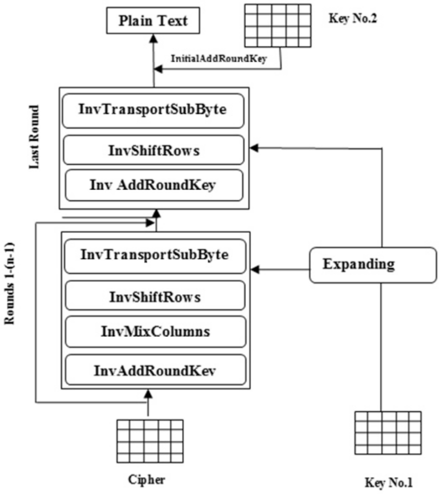


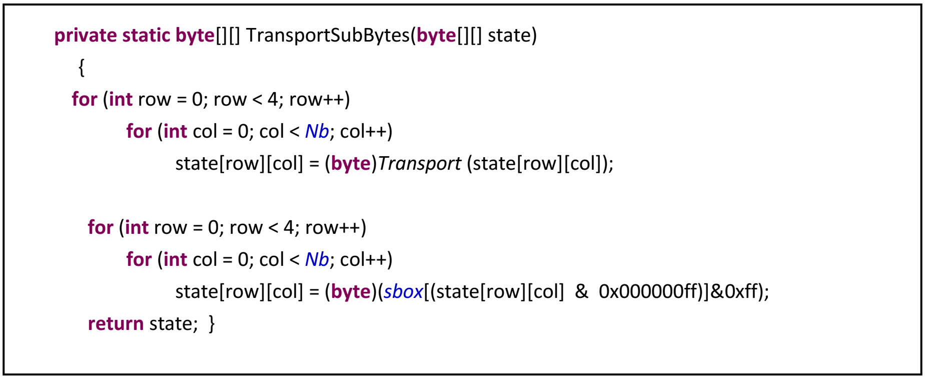
Fig. 2. Decryption Process of AES-R

In our research, we made the following three modifications to the original AES algorithm.

Adding Additional or Second key: In the proposed AES-R algorithm, operations are performed as shown in above figures and additional key (key No.2) will be added. The additional or second key length can reach up to 2048 bits.

Xoring second key with plain text: Before we perform the key expansion step of encryption process, the additional key will be first XORed with plain text. This XOR operation is called *InitialAddRoundKey*. The new output that resulted from *InitialAddRoundKey* operation is used as plain text for the following steps. After that the traditional key (key No.1) is expended to generate the sub-keys.

Modification in the SubBytes function: Instead of original SubBytes operation, we added a new operation called *Transport* in the original SubBytes operation. And we rename this SubBytes operation as *TransportSubBytes*. In the *TransportSubBytes* operation, data are transported first before they are substituted with S-Box values. Every element of the State array (i.e. 8 bits value) is divided into two halves (4 bits each) and these two halves are transported or swapped during the Transport process to get new State value. The java code of *TransportSubBytes* is shown in Fig. 3.



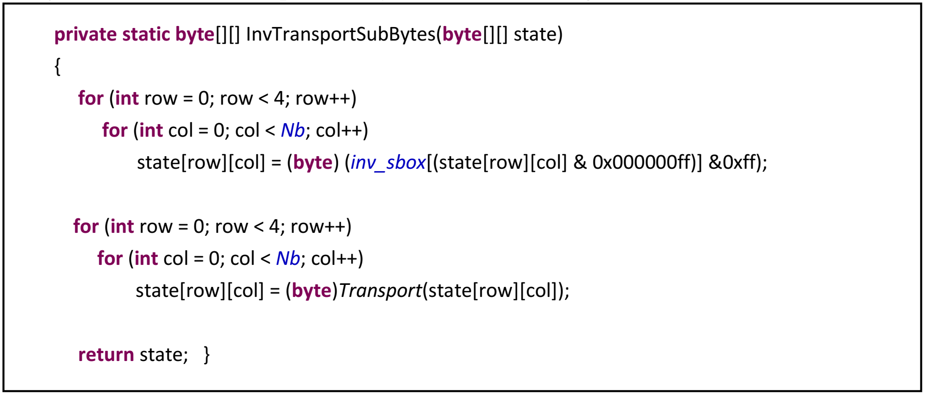
```
private static byte[][] TransportSubBytes(byte[][] state)
  {
 for (int row = 0; row < 4; row++)
        for (int col = 0; col < Nb; col++)
              state[row][col] = (byte)Transport (state[row][col]);

   for (int row = 0; row < 4; row++)
        for (int col = 0; col < Nb; col++)
              state[row][col] = (byte)(sbox[(state[row][col]  &  0x000000ff)]&0xff);
   return state;  }
```

Fig. 3. Java code of *TransportSubBytes* operation

The same process occurs in decryption operation of AES-R algorithm but by inversing the encryption processes (*InvTrasnportSubBytes*, *InvShiftRows* and *InvMixCoulmns*). Finally, we perform the *InitialAddRoundKey* operation by XORing additional key and the result of previous operations in order to retrieve back the original plain text.

The java code of *InvTransportSubBytes* is shown in Fig. 4.



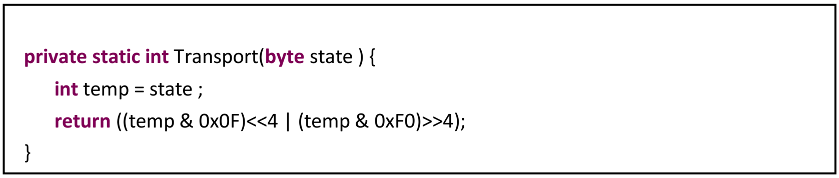
```
private static byte[][] InvTransportSubBytes(byte[][] state)
{
   for (int row = 0; row < 4; row++)
      for (int col = 0; col < Nb; col++)
           state[row][col] = (byte) (inv_sbox[(state[row][col] & 0x000000ff)] &0xff);

  for (int row = 0; row < 4; row++)
     for (int col = 0; col < Nb; col++)
          state[row][col] = (byte)Transport(state[row][col]);

   return state;   }
```

Fig. 4. Java code of *InvTransportSubBytes* operation.

The java code of *Transport* operation is shown in Fig. 5.



```
private static int Transport(byte state ) {
    int temp = state ;
    return ((temp & 0x0F)<<4 | (temp & 0xF0)>>4);
}
```

Fig. 5. Java code of *Transport* function

5 Experimental Result

For our experiment, we used a laptop with Windows 8.1 Pro, Intel® Core i5-4200U CPU @ 1.60 GHz Processor and 4 GB Memory.

Results of some experiments are given to prove its efficiency of application to text files. We used different sizes of text files to test the implementation of encryption and decryption processes. We used java programming language to implement two algorithms (traditional AES and AES-R). The performance is analyzed by calculating the execution time(in millisecond) for encryption and decryption. We used 128 bit of traditional key (key1) and 128 bit of additional key for performance analysis. Execution times of two algorithms are shown in tables to do comparison. The results are also shown as graphical drawing for easier observing. The Table 1 and Fig. 6 show the result of encryption operation.

Table 1., Executing Time to Different Files in millisecond (Encryption)

Test File		AES	AES-R
Test 1	(1 Kilo File)	1.371	0.970
Test 2	(2 Kilo File)	2.321	1.811
Test 3	(5 Kilo File)	5.418	4.151
Test 4	(10 Kilo File)	9.377	7.526
Test 5	(20 Kilo File)	18.351	15.088
Test 6	(30 Kilo File)	27.672	22.669
Test 7	(50 Kilo File)	44.568	32.273
Test 8	(75 Kilo File)	65.477	52.641
Test 9	(100 Kilo File)	103.011	65.927
Test 10	(1 Mega File)	867.852	717.192

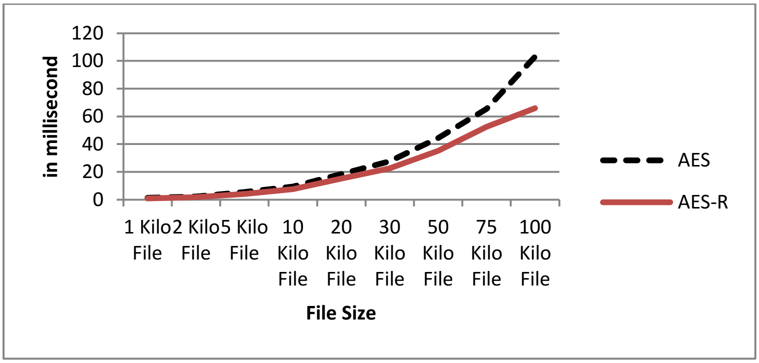


Fig. 6. Encryption Tests from AES and AES-R

Table 2 and Fig. 7 show the execution time of decryption in millisecond.

Table 2. Executing Time to Different Files in millisecond (Decryption)

Test File		AES	AES-R
Test 1	(1 Kilo File)	1.917	1.744
Test 2	(2 Kilo File)	3.281	3.164
Test 3	(5 Kilo File)	9.024	8.178
Test 4	(10 Kilo File)	15.915	15.421
Test 5	(20 Kilo File)	32.252	31.903
Test 6	(30 Kilo File)	47.903	46.916
Test 7	(50 Kilo File)	77.595	77.240
Test 8	(75 Kilo File)	116.076	115.392
Test 9	(100 Kilo File)	186.433	126.285
Test 10	(1 Mega File)	1540.999	1540.755

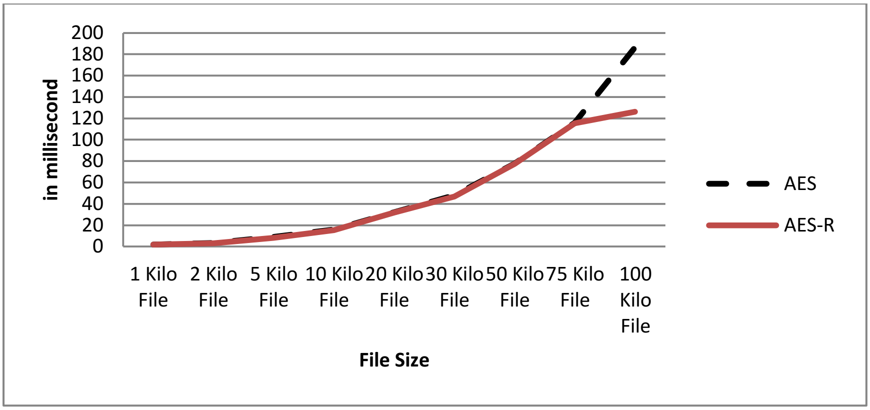


Fig. 7. Decryption Test results of AES and AES-R algorithms

5.1 Security Analysis

Avalanche effect. The avalanche effect is a desirable property of cryptographic algorithms. It is especially desirable when we design the cryptographic block ciphers algorithms and hash functions. In other words, the avalanche effect means significant changes in output when an input is changed slightly. Even a flip of single bit may lead to half or all of the output bits flip [13]. The Avalanche effect [14] can be calculated using

$$Avalanche\ effect = \frac{Number\ of\ flipped\ bits\ in\ cipher\ text}{Number\ of\ bits\ in\ Cipher\ text}. \quad (1)$$

If a block cipher does not exhibit the avalanche effect to a significant degree, then it has poor randomization, and thus a cryptanalyst can make predictions about the input, being given only the output. This may be sufficient completely or partially break the algorithm [8]. We tested the original algorithm and modified AES-R algorithm with 100 samples by changing the plain text by one bit. We found out that modified algorithm also produces the same result as original algorithm for one-bit change test. Table

3 shows some extracts of test results. Test results indicate that one bit change in plain text resulted in half of the output bits flips in both algorithms.

Table 3. , Results of plain text one-bit change test of AES and AES-R algorithm

Plain Text	Cipher Text	Algorithm
0xB9EE4700CCF1D4790	C970690B6C1D131DDDB61D7A0910DDD6 D222DFF34B61FFB21EADC2DFB2C4D5FF	AES
0xB9EE4700CCF1D4791	C970690B6C1D131DDDB61D7A0910DDD6 8E5CBC161901806DFE74E8B991083B58	
0xB9EE4700CCF1D4792	C970690B6C1D131DDDB61D7A0910DDD6 110BCF7FF1CA8DBF97F6E595E83E844B	
0xB9EE4700CCF1D4793	C970690B6C1D131DDDB61D7A0910DDD6 244E503E802D219DF1CAB02DCE69C558	
0xB9EE4700CCF1D4794	C970690B6C1D131DDDB61D7A0910DDD6 B3355F6E4B317D80AB75100AAA1FD453	
0xB9EE4700CCF1D4790	87E55CA9E5094186BD36B956FD306B7A C760B13F327678F9E3379C4A49846717	AES-R
0xB9EE4700CCF1D4791	87E55CA9E5094186BD36B956FD306B7A 452C82A99203912D67A1CDE993977164	
0xB9EE4700CCF1D4792	87E55CA9E5094186BD36B956FD306B7A 43BEC124DC59DC8C73F298E0767E6ABF	
0xB9EE4700CCF1D4793	87E55CA9E5094186BD36B956FD306B7A 10975F12573354FEDAA1B7D19EFB5E9B	
0xB9EE4700CCF1D4794	87E55CA9E5094186BD36B956FD306B7A 581E1F74001843107959F73415D9EFAA	

Again, we tried the one-bit change test by changing the encryption key1 with our AES-R algorithm. We used 100 samples to draw a conclusion. Like the AES algorithm, our proposed solution has generated completely different cipher text by changing one-bit of the key.

Information Entropy. Information theory is the mathematical theory of data communication and storage founded in 1949 by C.E. Shannon. Modern information theory is concerned with error-correction, data compression, cryptography, communications systems, and related topics [4]. The entropy is computed in bits and the entropy H(X) of the cipher text is calculated using

$$H(X) = \sum_{i=1}^{n} p_i \; log_2 \, (\frac{1}{p_i}). \tag{2}$$

P_i in (2) represents the probability of the symbol i. If 2^8 symbols with equal probability are used to get the entropy H(X) of the cipher text by applying (2), the ideal entropy of encrypted messages should be equal to 8, corresponding to a truly random source. In real world, the entropy value of the practical information source is usually lower than the ideal because a random message generated by a practical information source is not often random. In most cases, breaking an encryption system is difficult to perform if the entropy value of the cipher algorithm is close to the ideal [12].

We calculated the entropy H(X) of the cipher text that resulted from the original AES and AES-R using the same input file. The results are shown in Table 4.

Table 4. , Entropy of AES and AES-R for different files

Test File		AES	AES-R
Test 1	(1 Kilo File)	7.751	7.749
Test 2	(2 Kilo File)	7.845	7.861
Test 3	(5 Kilo File)	7.909	7.924
Test 4	(10 Kilo File)	7.935	7.939
Test 5	(20 Kilo File)	7.947	7.946
Test 6	(30 Kilo File)	7.934	7.939
Test 7	(50 Kilo File)	7.952	7.952
Test 8	(75 Kilo File)	7.952	7.950
Test 9	(100 Kilo File)	7.953	7.953
Test 10	(1 Mega File)	7.954	7.954

When the result table shown above was compared with the AES algorithm, the proposed AES-R has same encryption quality like AES and it is secure against entropy-based attacks similar to the original one.

6 Conclusion

After the AES has chosen as a standard symmetric cipher algorithm, assessment is done every five years to discover any attacks earlier and to try to make the algorithm stronger to stand against these attacks and make it active as long as possible. So the objective of assessment is to continually make improvements to make AES stronger than the attacks now and for future. The analysis results prove that the AES-R algorithm can be used as an option of AES when we would like to secure sensitive information. The proposed algorithm can also achieve the performance close to the original AES algorithm.

Both sender and receiver must use the same key in symmetric cipher algorithms. The other cipher algorithms use different keys but these keys must be related. In the AES-R algorithm we propose, the two keys do not need to be related at all. Our propose solution modified the AES algorithm by adding second secret key and modification at SubBytes operation. By using the additional secret key, it became more difficult to break the algorithm and it will take more time trying to break it. For that reason, AES-R can be consider as a good choice for some applications which require confidentiality to protect sensitive data because the time taken to encrypt and decrypt data is close to or sometimes even better than AES algorithm in encryption.

As for future work, the proposed AES-R algorithm will be tested with other testing methods to analyze the security performance for the evaluation purpose. In addition to testing with text data, we will apply the same algorithm on images and audio data in further research. Another research could be using image data as a watermark and an additional secret key and evaluate the security performance and the time taken of that modification.

References

1. Ashraf Odeh, Shadi R.Masadeh, Ahmed Azzazi, "A performance evaluation of common encryption techniques with secure watermark system(SWS)", International Journal of Network Security & Its Applications(IJNSA), vol. 7, No. 3, pp. 31-38, 2015.
2. Ibtihal Mohamed Abdullateef Fadul, Tariq Mohamed Hassan Ahmed, "Enhanced security of Rijndael algorithm using two secret keys", International Journal of Security and Its Applications, vol. 7, no. 4, pp. 127-134, 2013.
3. Chittaranjan Pradhan and Ajay Kumar Bisoi, "Chaotic variations of AES algorithm", International Journal of Chaos, Control, Modeling and Simulation (IJCCMS), vol.2, no.2, pp. 19-25, 2013.
4. Reena Mehla and Harleen Kaur, "Different reviews and variants of advance encryption standard", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 Impact Factor (2012):3.358, pp. 1895-1896., 2012
5. Abdulkarim Amer Shtewi, Bahaa Eldin M. Hasan, Abd El Fatah .A. Hegazy, "An efficient modified advanced encryption standard (AES-R) adapted for image cryptosystems", International Journal of Computer Science and Network Security, vol. 10, no. 2, pp. 226-232, 2010.
6. Kazys Kazlauskas, Robertas Smaliukas and Gytis Vaicekauskas, "A novel method to design S-Boxes based on key-dependent permutation schemes and its quality analysis", International Journal of Advanced Computer Science and Applications, vol. 7, No. 4, 2016.
7. Sumira Hameed, Faisal Riaz ,Riaz Moghal, Gulraiz Akhtar, Anil Ahmed and Abdul Ghafoor Dar, "Modified advanced encryption standard for text and images", Computer Science Journal, vol. 1, Issue 3, pp. 120-129, 2011.
8. Krishnamurthy G N, V Ramaswamy, "Making AES stronger: AES with key dependent S-Box", IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 9, pp. 388-38, 2008.
9. Obaida Mohammad Awad AL-Hazaimeh, "A new approach for complex encryption and decryption Data", IJCNC International Journal of Computer Network and Communications, vol. 5, no.2, pp. 95-103, 2013.
10. Sliman Arrag, Abdellatif Hamdoun, Abderrahim Tragha, Salah Eddine Khamlich, "Implementation of stronger AES by using dynamic S-Box dependent of master Key", Journal of Theoretical and Applied Information Technology, vol. 8, no. 9, pp. 196-204, July 2013, ISSN: 1992-8645, E-ISSN: 1817-3195.
11. Ali Abdulgader, Mahamod Ismail, Nasharuddin Zainal, Tarik Idbeaa, "Enhancement of AES algorithm based on chaotic maps and shift operation for image encryption", Journal of Theoretical and Applied Information Technology, vol. 71, no. 1, pp. 1-12, January 2015, ISSN: 1992-8645, E-ISSN: 1817-3195.
12. G.Sai Akhil., M. Amarndah, Kiran Kumar, "A Technique to Secure Data Storage and Sharing Using AES", International Journal of Scientific & Engineering Research Volume 7, Issue 12, pp. 35-39, December-2016, ISSN 2229-5518
13. Maqsood Mahmud, Muhammad Khurram Khan, Khaled Alghathbar, "Biometric-Gaussian-Stream (BGS) Cipher with new Aspect of Image Encryption (Data Hiding)", BSBT 2009, CCIS 57, Springer-Verlag Berlin Heidelberg, pg. 98, 2009
14. Shraddha Dadhich, "Performance analysis of AES and DES cryptographic algorithms on windows & ubuntu using java", International Journal of Computer Trends and Technology (IJCTT), vol. 35, no.4, pp. 179-183, May 2016, ISSN: 2231-2803.

Residential Neighbourhood Security using WiFi

Kain Hoe Tai[1], Vik Tor Goh[1](✉), Timothy Tzen Vun Yap[2], and Hu Ng[2]

[1] Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
[2] Faculty of Computing & Informatics, Multimedia University, Cyberjaya, Malaysia
kainhoe@gmail.com,
{vtgoh, timothy, nghu}@mmu.edu.my

Abstract. This paper focuses on the design of a WiFi-based tracking and monitoring system that can detect people's movements in a residential neighbourhood. The proposed system uses WiFi access points as scanners that detect signals transmitted by the WiFi-enabled smartphones that are carried by most people. Our proposed system is able to track these people as they move through the neighbourhood. We implement our WiFi-based tracking system in a prototype and demonstrate that it is able to detect all WiFi devices in the vicinity of the scanners. We describe the implementation details of our system as well as discuss some of the results that we obtained.

Keywords: WiFi · residential security · tracking and monitoring

1 Introduction

Property crimes often result in financial loss, psychological trauma and even deflation in property value. The Royal Malaysian Police reported 142,000 cases in 2017 alone for property crimes that include burglary, arson, vandalism and vehicle theft [18]. Currently, responses to these incidents have been reactive, namely relying on the ability of the police force and legal fraternity to investigate and prosecute. Preventive measures such as better home security and constant surveillance have been effective but limited due to cost. Only homeowners or neighbourhoods that can afford to pay for surveillance equipment (e.g. CCTV, motion sensors, etc.) or gated/guarded facilities will be better protected. Communities that are less financially able are therefore more susceptible to such crimes.

In this project, we propose a cost-effective surveillance system that can provide around-the-clock monitoring without the need for extra equipment and associated costs. The proposed system relies on the 802.11 wireless fidelity or WiFi protocol [14]. This is possible because of the prevalence of portable WiFi devices, namely smartphones and the ubiquity of wireless fidelity networks or WiFi networks in most residential premises. According to the Internet Users Survey 2017 that was conducted by MCMC [8], 68.7% of the 32 million people in Malaysia (citizen or otherwise) own smartphones. Furthermore, the report also states that there are about 2.5 million fixed broadband subscribers in Malaysia. These broadband packages often include WiFi connectivity.

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_43

Our proposed surveillance system aggregates these existing WiFi networks in a residential neighbourhood into a virtual neighbourhood surveillance zone that can track and monitor the comings and goings of individuals. Within the surveillance zone, our system then monitors the 802.11 wireless traffic that is emitted by WiFi-enabled smartphones to observe peoples movements.

The proposed system is intended to be a complementary surveillance system (as opposed to a substitute) that provides constant monitoring in a neighbourhood without any additional costs. It enhances security, improves neighbourhood safety, and deters potential property crimes.

2 Related Work

Modern home security solutions typically consist of vibration sensors, motion detectors (e.g. passive infrared motion detector), magnetic switches, and even closed circuit television (CCTV). By and large, these sensors have remained the same for many years because of their simplicity and convenience. Despite that, there are some novel home security sensors such as those by [7]. They proposed to use stepped-FM ultra-wideband (UWB) sensors to detect intruders. Unlike most typical sensors, this sensor can determine relative distance between potential intruders and break-in points as well as estimate the intrusion port (e.g. doors or windows). However, this sensor has difficulties in discerning human identities. When activated, this system will identify everyone that approaches the surveillance zone as intruders, even if the detected person is in fact the owner of the residence.

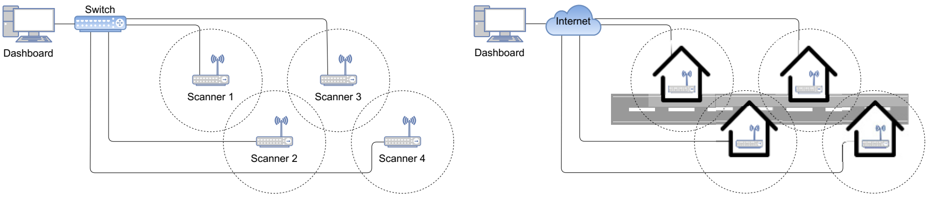
To overcome the lack of computational or "thinking" capabilities, many home security solutions are being integrated into smart homes. A smart home utilises home automation technologies to control lighting, climate, appliances, and home security systems. For example, [1], [6], and [16] have proposed intelligent systems that use computational algorithms to control the operations of the home security systems. Although these algorithms increase the efficiency, capabilities, and accuracy of home security systems, they are costly to install and difficult to maintain due to their increased complexity.

Although WiFi-based positioning systems are already quite established, most of the current systems are designed for indoor positioning [2, 5, 13] or retail analytics [12, 15, 17]. Many of these systems focus on improving the accuracy of WiFi-based indoor positioning algorithms, which also include developing better ways to track people's movements in an indoor environment as well as predict shoppers' behaviour. To the best of our knowledge, there are no previous work that attempt to improve residential or neighbourhood security by using the WiFi protocol.

In our work, we aim to design a WiFi-based tracking system that can detect and monitor people through their Wifi-enabled smartphones. The proposed system is intended to be used outdoors so that it can monitor people as they move through the neighbourhood. This is especially useful when we need to track people of questionable intentions. Our proposed system uses RSSI localisation techniques due to their simplicity and ease of deployment but with the addition of handoff methods so that movements can be tracked across multiple access points.

3 Design and Implementation

The WiFi-based tracking system aims to monitor and follow a person's movement in the area under surveillance. In specific, the proposed system is envisioned to be used in a residential neighbourhood, as opposed to the more typical usage of WiFi as indoor positioning systems. In this work, we study the feasibility of such a system through the design and implementation of a small-scale prototype.



(a) Prototype's topology (b) Envisioned topology

Fig. 1. Topology of proposed WiFi tracking system.

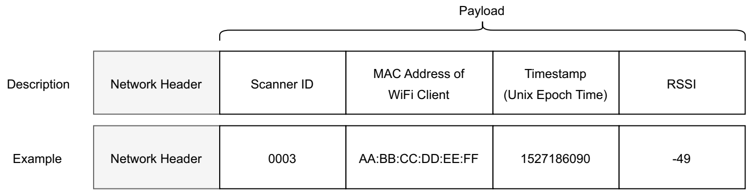
In our prototype, we utilise WiFi access points that have been configured into monitor mode. In monitor mode, these access points can monitor all WiFi traffic transmitted in the wireless network, thus making them ideal *scanners* of other wireless devices. These scanners in our prototype system are then placed to mimic the usual placements of access points in a residential neighbourhood as shown in Fig. 1.

The overlapping signals allow the scanners to monitor an individual as he/she moves between them, thus providing better tracking capabilities.

3.1 Implementing the Scanners

As shown in Fig. 1a, the proposed tracking system consists of two components, namely the scanners and dashboard. In the prototype, we used TP-Link MR3020 wireless routers as the scanners. These devices were chosen because their small form factor made them easier to handle, they supported the Open-WRT custom router firmware, and most importantly, they were within the limited budget of the project. In our prototype system, the wireless routers are connected to the dashboard via a switch.

We use the Linux-based Open-WRT firmware because of its versatility which allows us to reconfigure the routers into monitor mode. This is vital because the proposed tracking system relies on a WiFi device's behaviour of carrying *active scans* to detect nearby access points. In active scanning, the device broadcasts probe request frames and listens to probe responses. As such, if the the monitor-mode routers detect these probe request frames, it can be concluded that a WiFi device is in the vicinity of that scanner. These probe request frames are unique to each WiFi device because the frames contain identifying information, namely the MAC address of the WiFi device [3].

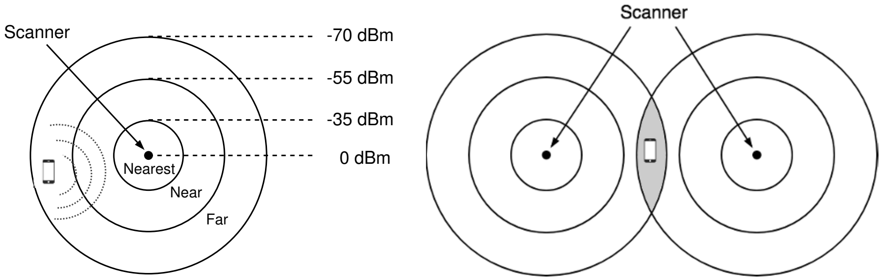


		Payload			
Description	Network Header	Scanner ID	MAC Address of WiFi Client	Timestamp (Unix Epoch Time)	RSSI
Example	Network Header	0003	AA:BB:CC:DD:EE:FF	1527186090	-49

Fig. 2. Packet structure containing information from scanner.

Whenever the scanners detect the presence of a WiFi device through the probe request frames, it first extract usable information such as device's MAC address. It also measures the Received Signal Strength Indicator (RSSI) from the device. These information are formatted as shown in Fig. 2 and then sent by the scanners to the dashboard for further processing.

3.2 Implementing the Dashboard



(a) Location regions

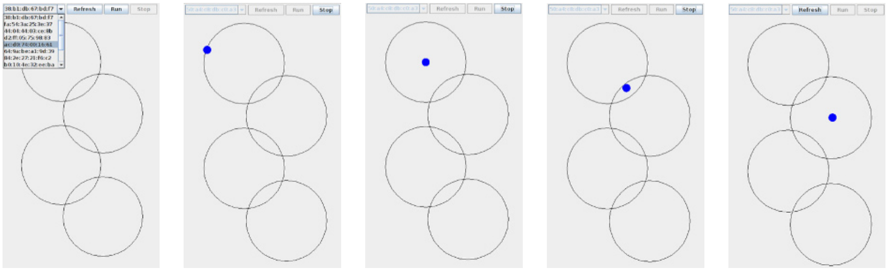
(b) Overlapping detection

Fig. 3.

As the motivation of the proposed system is to track a person's movements (as opposed to a positioning problem) through the residential neighbourhood, we do not need a high level of accuracy when determining the person's whereabouts. We need only have an estimate of the person's location relative to the scanners. The location can be estimated based on the correlation between RSSI levels and distance [9]. In general, larger (stronger) RSSI levels indicate that the WiFi transmitter is nearer to the receiver, while lower (weaker) RSSI levels indicate that the WiFi transmitter is farther away.

Fig. 3a shows how the signal reception area of the scanner is divided into three regions; nearest, near, and far. The WiFi device's location is determined to be in any one of these regions based on its RSSI level. We set the lower limit to -70 dBm because any signal reception below this tends to be unstable and unreliable. Besides that, if a WiFi device is detected by an adjacent scanner (as depicted in Fig. 1a), the device will be shown to be present in the overlapping region of both scanners as shown in Fig. 3b.

Although not as accurate traditional WiFi positioning methods, our approach is simple and straightforward, thus eliminating the need for complex trilateration calculations or pre-determined WiFi fingerprints as explained in [11]. More importantly, our approach is sufficient for the purpose of tracking a person's movements.



$t = 0$ $t = 1$ ... ... n

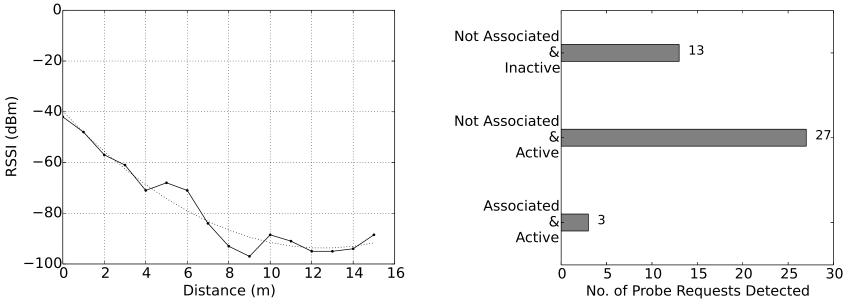
Fig. 4. Tracking a person's movement.

These localisation techniques for the dashboard are implemented in Java. The Java software processes the packets received from the scanners and then visualises the location of the WiFi device as shown in Fig. 4. A list of detected WiFi devices are shown in the dropdown menu for the user to choose from. Once the user has chosen the WiFi device, the system will continuously track that device as it moves across the scanners. As it can be clearly seen, the proposed system is able to track the WiFi device as it moves.

4 Discussion and Analysis

4.1 RSSI and Distance

We performed an experiment to measure the relationship between RSSI and distance in a WiFi network. The experiment was carried out with the scanner placed in a residential premise while the WiFi device was slowly moved farther away from it. The results of this experiment are shown in Fig.5a. As expected, RSSI becomes weaker as the WiFi device moves farther away from the scanner.



(a) RSSI vs. Distance (b) Probe requests in 1 minute

Fig. 5. Results and observations of WiFi protocol.

An interesting observation is that after 4 meters, the RSSI fluctuates more, between -70 to -100 dBm. This could be due to multipath propagation which becomes more prevalent as the distance between scanner and device increases. On the other hand, when the device is closer to the scanner, the path from the device to the scanner is more direct and therefore, less susceptible to the effects of multipath propagation.

In Section 3.2, it was stated that the lower limit was set to -70 dBm because any signal received below this level tends to be unstable and unreliable. If we observe Fig. 5a, −70 dBm corresponds to a distance of 6 meters between scanner and the WiFi device. This short distance of the prototype system is because TP-Link MR3020 was designed to a portable device, only intended for personal use with a range of about 10 meters only. In comparison, traditional home access points have ranges between 30 – 70 meters [10].

4.2 Analysing the Frequency of Probe Requests

Although yet to be implemented in the current iteration of our prototype, one of the design objectives is for the system to be able to distinguish between WiFi devices that belong to residents and devices that belong to strangers. To that end, we carried out another experiment to examine the 802.11 association process. Specifically, we want to study the frequency of probe requests and the possibility of using it to differentiate between residents and strangers.

There are two states in the association process; associated or unassociated. These states are described in [4]. There are another two states in the WiFi device's status; active or inactive. An active device is being actively used while an inactive device is on standby and not being used. In this experiment, we measure the number of probe requests sent by the following states:

1. Not Associated and Inactive
2. Not Associated and Active
3. Associated and Active

We measure the number of probe requests sent in 1 minute by a WiFi device that has been placed approximately 2 meters away from the scanner. The results of the experiment are shown in Fig. 5b. In the figure, we can quite easily differentiate between the "not associated" and "associated" WiFi devices. Even the "not associated and inactive" state has four times the number of probe requests compared to the "associated and active" state. This is a useful feature because most strangers to a residential neighbourhood are not associated to a WiFi network, thus allowing the proposed system to only track persons of interest. We intend to incorporate this information into our future versions of the prototype.

5 Conclusion and Future Work

In this project, we aimed to develop a novel and cost-effective monitoring system to improve the security of a residential neighbourhood. We achieved this by taking advantage of the ubiquity of WiFi networks that can be found in most residential premises as well as the proliferation of WiFi-enabled smartphones. Due to the results of the MCMC survey, we assume that most people coming into the neighbourhood will own a smartphone and therefore, can be "seen" and uniquely identified. In order to "see" these people, we configure WiFi access points to detect probe requests that are constantly transmitted by the smartphones. Once detected, our proposed system will then track and monitor persons of interest as they move through the neighbourhood.

For our future work, we intend to include some differentiation algorithms to distinguish between residents and non-residents. This feature will allow users to isolate persons of interest from the usual residents, thus improving usability. Besides that, we also plan on scaling up the project over a larger geographical area by utilising typical access points with greater ranges instead of the shorter-range portable access points.

In conclusion, we have implemented the proposed system as a small-scale prototype and have demonstrated that it functions as expected. The Java-based user interface is able to process data from the scanners and track people's movements as they move between different scanners.

Acknowledgements. Financial support from the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme with grant number FRGS/1/2015/SG07/MMU/02/1, as well as the Multimedia University Capex Fund with Project ID MMUI/CAPEX170008, are gratefully acknowledged.

References

1. Ahmad, A.W., Jan, N., Iqbal, S., Lee, C.: Implementation of zigbee-gsm based home security monitoring and remote control system. In: Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on. pp. 1–4. IEEE (2011)

2. Bose, A., Foh, C.H.: A practical path loss model for indoor wifi positioning enhancement. In: Information, Communications & Signal Processing, 2007 6th International Conference on. pp. 1–5. IEEE (2007)
3. Corbett, C.L., Beyah, R.A., Copeland, J.A., et al.: Using active scanning to identify wireless nics. In: Proceedings of IEEE Information Assurance Workshop (IAW) (2006)
4. Dionicio, R.: 802.11 state machine – association and authentication (Nov 2015), `https://www.packet6.com/802-11-state-machine/`
5. Evennou, F., Marx, F.: Advanced integration of wifi and inertial navigation systems for indoor mobile positioning. Eurasip journal on applied signal processing **2006**, 164–164 (2006)
6. Hou, J., Wu, C., Yuan, Z., Tan, J., Wang, Q., Zhou, Y.: Research of intelligent home security surveillance system based on zigbee. In: Intelligent Information Technology Application Workshops, 2008. IITAW'08. International Symposium on. pp. 554–557. IEEE (2008)
7. Jitsui, Y., Kajiwara, A.: Home security monitoring based stepped-fm uwb. In: Antenna Technology (iWAT), 2016 International Workshop on. pp. 189–191. IEEE (2016)
8. Malaysian Communications and Multimedia Commission: Internet Users Survey 2017 (2017)
9. Mazuelas, S., Bahillo, A., Lorenzo, R.M., Fernandez, P., Lago, F.A., Garcia, E., Blas, J., Abril, E.J.: Robust indoor positioning provided by real-time rssi values in unmodified wlan networks. IEEE Journal of selected topics in signal processing **3**(5), 821–831 (2009)
10. Mitchell, B.: How far will your wifi reach? (Feb 2018), `https://www.lifewire.com/range-of-typical-wifi-network-816564`
11. Mok, E., Retscher, G.: Location determination using wifi fingerprinting versus wifi trilateration. Journal of Location Based Services **1**(2), 145–159 (2007)
12. Prasertsung, P., Horanont, T.: How does coffee shop get crowded?: using wifi footprints to deliver insights into the success of promotion. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. pp. 421–426. ACM (2017)
13. Shin, B.J., Lee, K.W., Choi, S.H., Kim, J.Y., Lee, W.J., Kim, H.S.: Indoor wifi positioning system for android-based smartphone. In: Information and Communication Technology Convergence (ICTC), 2010 International Conference on. pp. 319–320. IEEE (2010)
14. Walrand, J., Parekh, S.: Communication networks: a concise introduction. Synthesis Lectures on Communication Networks **3**(1), 1–192 (2010)
15. Wang, Y., Yang, J., Liu, H., Chen, Y., Gruteser, M., Martin, R.P.: Measuring human queues using wifi signals. In: Proceedings of the 19th annual international conference on Mobile computing & networking. pp. 235–238. ACM (2013)
16. Ye, X., Huang, J.: A framework for cloud-based smart home. In: Computer science and network technology (ICCSNT), 2011 international conference on. vol. 2, pp. 894–897. IEEE (2011)
17. Zeng, Y., Pathak, P.H., Mohapatra, P.: Analyzing shopper's behavior through wifi signals. In: Proceedings of the 2nd workshop on Workshop on Physical Analytics. pp. 13–18. ACM (2015)
18. Zolkepli, F., Camoens, A.: Igp: Crime index down by 11.7 percent (Mar 2018), `https://www.thestar.com.my/news/nation/2018/03/25/igp-crime-index-down-by-11_7-percent/`

Prediction of Mobile Phone Dependence Using Bayesian Networks

Euihyun Jung

Dept. of Convergence Software, Anyang University, Anyang City, Korea
jung@anyang.ac.kr

Abstract. Bayesian Networks have been widely used in various domains, but they have been rarely used in educational domain. In this paper, we discover a Bayesian Network model to figure out variables related to adolescents' mobile phone dependence and their influences. For this study, Markov Blanket is used to identify the strongly related variables with the Korea Children and Youth Panel Survey (KCYPS) data. From the analysis with the discovered BN, "attention ability", "depression", "caregiver's abuse", "fandom activity", and "aggression" are extracted as the variables related to adolescents' mobile phone dependence. These results suggest that considering the variables and their interactions are useful to adjust adolescent's mobile phone dependence. This paper also shows Bayesian Networks are adequate to find the interdependence of variables and their causal relationships in educational domain.

Keywords: Bayesian Network, Markov Blanket, Data Mining, Mobile Phone Dependence.

1 Introduction

Bayesian Networks (BNs) have been successfully applied in various areas such as medicine, biology, health, finance, etc. [1][2][3][4]. A BN is a graphical representation of joint probability distributions of variables based on probability theory. It is a directed acyclic graph (DAG) in which nodes represent variables and arcs describe probabilistic dependencies between the nodes [5]. BNs are useful to find out the relationship of variables because the structure of the associated DAG represents the dependencies among variables and gives a concise specification of joint probability distributions. BNs allow researchers to make inferences efficiently even when they have many variables in the network [6][7][8]. Once a BN model has been gotten, it can be used to figure out the complex interactions and causal relationships among variables.

Due to their advantages, BNs have been widely used in various domains, but they are rarely applied in educational domain. The majority of researchers in educational area have still used the traditional statistic methods because they are unfamiliar with BNs and most existing statistical packages don't deal with BNs. The statistical approach is basically good for most situations, but researchers may miss important fea-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_44

tures because they cannot cover all cases. Comparing to this, the Bayesian Networks approach in this paper does not make a hypothesis but just discover a BN from the data instead. Once the BN is discovered, it used to figure out the related variables and explain the degrees of effects.

Mobile phone use has increased in recent years. According to the information provided research of Korea Internet & Security Agency [9], approximately 89.5% of the inhabitants in Korea owned a mobile phone in 2017. Specially, there has been an increasing trend of mobile phone use among students [10]. Mobile phone dependence refers to excessive use and mobile phone use in public places even when such use is considered to be a nuisance [11]. According to the research of Ministry of Science and ICT and National Information Society in Korea [12], the rate of adolescents' mobile phone dependence in 2016 was 30.6% that is the highest of all ages. Due to its negative physical and psychological consequences of the excessive use of mobile phones [10], the study finding factors related to mobile phone dependence has become important in order to relieve the dependence [13]. However, an empirical study on mobile phone dependence among adolescents has not been widely performed yet [14][15]. In this paper, we suggest how to find the related variables to mobile phone dependence and the effects of the variables with BNs.

The rest of this paper is organized as follows. Section 2 describes the data set used in the analysis and a discovered BN with Markov Blanket. In Section 3, several analyses with the discovered BN are conducted and Section 4 concludes the paper.

2 Discovery of A Bayesian Network Model

2.1 Overview of Dataset

In this study, the Korea Children and Youth Panel Survey (KCYPS) longitudinal data was used. The data was collected by the National Youth Policy Institute (NYPI) of South Korea [16]. Researchers used the stratified multistage clustering for the KCYPS sampling and collected data through seven follow-up surveys from 2010 to 2016. The KCYPS data represent Korean teens and children. At the time of the first study (2010), the subjects were in the 7th grade classes. Among the KCPYS data, the second wave (2011) was chosen in order to analyze adolescents' mobile phone dependence and they were in the 8th grade classes.

In general, missing values and noises can be included in the longitudinal data. Therefore, data preprocessing is necessary to improve the quality of the data. Each variable should be also simplified in order to learn a BN. In this study, variables with a 4-point Likert scale were categorized as binary scales (Yes/No) and variables with a 5-point Likert scale were turned into three-level scale (High/Average/ Low). A total of 1,792 subjects were included after omitting missing values. The 43 variables are summarized in Table 1.

Table 1. Description of variables in the data set

Variable	State	Meaning
Dad's educational background	M	Middle school
	H	High school
	C	College
	U	University
	G	Graduate School
Health, School friends, School teachers	Y	Good
	N	Bad
Language grade, Mathematics grade English grade, Science grade, Society grade	H	High
	A	Average
	L	Low
Grade satisfaction, Social withdrawal, Aggressive, Attention, Depression, Smoking, Drinking, Absence, Runaway home, Taunting, Bullying, Assaulting, Gang fight, Threating, Taking away, Stealing, Taunted, Bullied, Assaulted, Threatened, Taken away, Neglect, Abuse, Mobile phone dependence, Parents know my friends, Parents meet my friends, Parents like my friends, Computer game act, Fandom act	Y	Yes
	N	No
Sense of community, Multi-culture attitudes, Sense of local community	Y	Have
	N	Don't have
School act	Y	Do
	N	Don't do
School rule	Y	Follow
	N	Don't follow

2.2 The Discovered Bayesian Network Model

In this paper, a statistical computing tool, R is used to discover a BN model. Hill-climbing method [18] is used as a learning algorithm to construct a BN model. Figure 1. (a) shows the part of the initial BN model. Then, in order to find variables strongly related to mobile phone dependence, Markov Blanket of phone variable is investigated. The Markov Blanket of a node includes its parents, its children, and the children's other parents [17]. The extracted Markov Blanket from the discovered BN model is shown in Figure 1. (b).

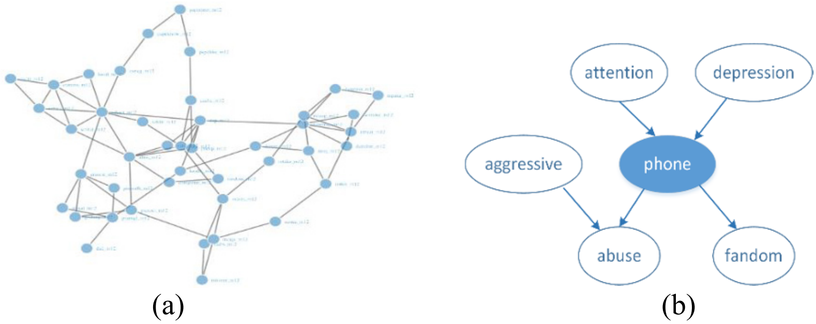


Fig. 1. (a) The initial discovered BN model with the Hill-climbing (hc) learning method. (b) The extracted variables related to Mobile Phone Dependence with Markov Blanket.

From the Markov Blanket of *phone* variable, *attention*, *depression*, *abuse*, and *fandom* variables are directly related. Besides, *aggressive* variable is indirectly related. The results of the Markov Blanket are summarized in Table 2.

Table 2. The strongly related variables to 'Mobile Phone Dependence'.

Name	Description	Value	Relation
attention	Attention ability	1-yes, 2-no. Yes means I have good attention.	Direct
depression	Depression	1-yes, 2-no. Yes means I feel like depression.	Direct
abuse	Caregiver's abuse	1-yes, 2-no. Yes means I have experience.	Direct
fandom	Fandom activity	1-yes, 2-no. Yes means I do fandom activity.	Direct
aggressive	Aggression	1-yes, 2-no. Yes means I'm aggressive.	Indirect

3 Analysis

3.1 Interpretation of the BN Model

While the bnlearn package in R allows us to discover a BN model, implementing the BN in Netica [19] enables us to simulate the effects by changing the value of variables' probabilities. Netica is the one of commercial software tools that automate the process of inference and the results of the inference are shown graphically using bar charts [5]. The initial BN model with the probabilities in Netica is shown as Figure 2. For instance, the 'yes' value of 15.2 in the depression node means an interviewee undergoes depression with the probability P = .152.

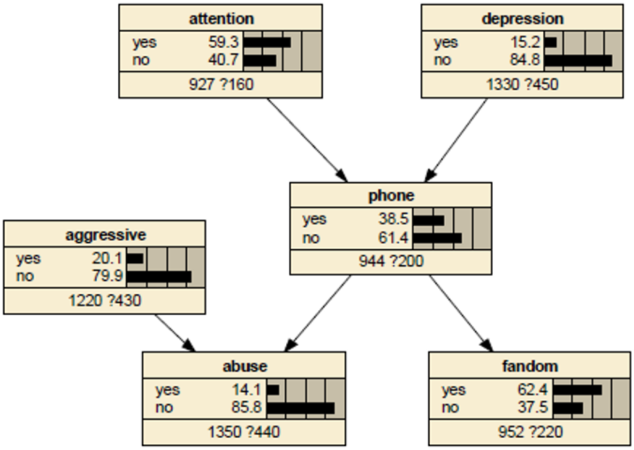


Fig. 2. The initial BN model with the probabilities for variables.

3.2 Changing Probability Values

Reasoning from Changes of Explanatory Variables. From the BN implemented in Netica, it is possible to predict the effects by changing the probabilities of explanatory variables. Table 3 provides the four cases of maximizing the probabilities of the explanatory variables. Figure 3 show the case of maximizing the 'no' value of *attention* and the 'yes' value of *depression* variable.

Table 3. Changes of explanatory variables' probabilities and their effects

attention	*depression*	*phone* (yes)	*phone* (no)
Yes=59.3, No=40.7 (initial values)	Yes=15.2, No=84.8	38.5	61.4
Yes=100.0	Yes=100.0	53.6	46.4
Yes=100.0	**No=100.0**	27.6	**72.4**
No=100.0	**Yes=100.0**	**65.7**	34.3
No=100.0	No=100.0	45.8	54.2

When we maximize the 'no' value of attention variable from 40.7% to 100.0% and the 'yes' value of depression variable from 15.2% to 100.0%, the 'yes' value of phone variable goes up from 38.5% to 65.7% as shown in Fig. 3. This is the highest value of phone variable when we change the probabilities of the explanatory variables. It indicates that youths who have bad attention and feel depression may depend on mobile phone with a high probability. On the other hand, when the explanatory variables have the opposite values, the 'no' value of phone rises from 61.4% to 72.4%. It means that youths who have good attention and don't feel like depression are not likely to depend on the mobile phone.

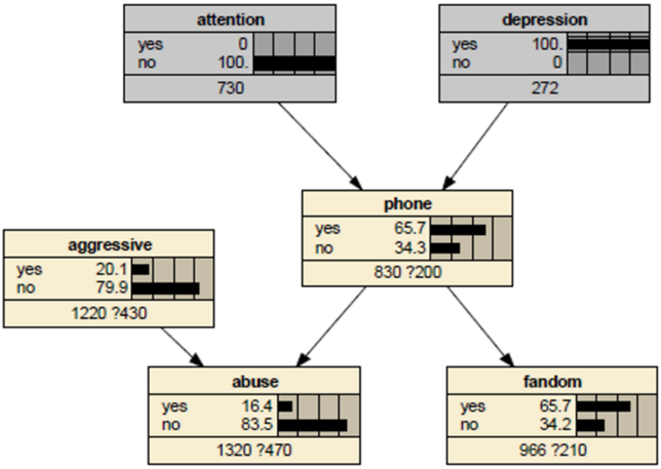


Fig. 3. The one case of maximizing the probabilities of the explanatory variables

Reasoning from Changes of the Dependent Variable. In BNs, it is also possible to identify the changes of the probabilities of the related variables by changing the dependent variable (phone variable). We maximized the dependent variable in two ways. One is the maximum of 'yes' value (=100.0) and the other is the maximum of 'no' value (=100.0). Their effects on the related variables are summarized in Table 4.

When the 'yes' value of *phone* variable is maximized, the probabilities of *attention* (no), *depression* (yes), *abuse* (yes), and *fandom* (yes) variables increase. In contrast, if the 'no' value of *phone* variable is maximized, the probabilities of *attention* (yes), *depression* (no), *abuse* (no), and *fandom* (no) variables increase. These results suggest that *phone* variable has a positive influence on the *depression*, *abuse*, and *fandom* variables and has a negative influence on *attention* variable. Meanwhile, *phone* variable doesn't affect the *aggressive* variable directly.

4 Conclusion

Although researchers in many domains have chosen BNs in order to produce an intuitive, transparent, and graphical representation of the investigated variables for identifying the interdependencies of variables and their causal relationships [17], there has been very little research using BNs in educational domain. However, if BNs are used in educational domain, they will provide useful insights in the context of education.

Mobile phones have become an essential part of modern life and especially there has been an increasing trend of mobile phone use among adolescents. Recently, adolescents' excessive use of mobile phones has now been recognized as a serious social problem. However, mobile phone dependence has not been widely studied using various methods. Therefore, it is difficult to suggest which variables are related to the mobile phone dependence and how much the variables affect the degree of the dependence.

In this paper, we try to find the variables explaining the mobile phone dependence and how the variables are connected to each other using the BN. We adopt the hill-climbing method in order to build a BN model and apply Markov Blanket to draw the variables strongly related to mobile phone dependence. The variables are "attention ability", "depression", "caregiver's abuse", "fandom activity", and "aggression".

By reasoning from changes of explanatory variables, we conclude that "attention ability" and "depression" have stronger effects on mobile phone dependence than other variables. We also observe the change of the dependent variable (mobile phone dependence) affect the probabilities of "attention ability", "depression", "caregiver's abuse", and "fandom activity". That is, "attention ability" is negatively but the other three variables are positively related to the mobile phone dependence.

This finding of related variables and their effects can be used to develop educational programs which help to adjust adolescents' mobile phone dependence. In the future, we are going to investigate the interdependencies and causal-effects relationships of other variables using various BN methods in order to provide useful results in educational domain.

Table 4. Changes of the dependent variable' probabilities and their effects

Dependent Variable	Related Variables		Prior Probability	Posterior Probability
phone: Yes=100	*attention*	Yes	59.3	48.4
		No	40.7	**51.5**
	depression	Yes	15.2	**23.0**
		No	84.8	76.9
	abuse	Yes	14.1	**19.4**
		No	85.8	80.6
	fandom	Yes	62.4	**70.0**
		No	37.5	30.0
	aggressive	Yes	20.1	20.1
		No	79.9	79.9
phone: No=100	*attention*	Yes	59.3	**66.0**
		No	40.7	33.9
	depression	Yes	15.2	10.2
		No	84.8	**89.7**
	abuse	Yes	14.1	10.8
		No	85.8	**89.1**
	fandom	Yes	62.4	57.7
		No	37.5	**42.3**
	aggressive	Yes	20.1	20.1
		No	79.9	79.9

References

1. Djebbari, A., & Quackenbush, J. (2008). Seeded Bayesian networks: Constructing genetic networks from microarray data. *BMC Systems Biology, 2008*, 2-57.
2. Lewis, F.I., & McCormick, B. J. (2012). Revealing the complexity of health determinants in resource-poor settings. *American Journal of Epidemiology, 176*(11), 1051-1059.
3. Lucas, P. (2004). Bayesian analysis, pattern analysis, and data mining in health care. *Current Opinion in Critical Care, 10*(5), 399-403.
4. Shenoy, C., Shenoy, P. P. (2000). *Bayesian network models of portfolio risk and return*. The MIT Press.
5. Nadkarni, S., & Shenoy, P. P. (2001). A Bayesian network approach to making inferences in causal maps. *European Journal of Operational Research, 128*, 479-498.
6. Aguilera, P. A., Fernandez, A., Fernandez, R., Rumi, R., & Salmeron, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software, 26*, 1376-1388.
7. Dlamini, W. M. (2010). A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. *Environmental Modelling & Software, 2*, 199-208.
8. Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.
9. Korea Internet & Security Agency (2018). *The survey of 2017 internet use*. Naju: Korea Internet & Security Agency (KISA).
10. Nikhita, C. S., Jadhav, P. R., & Ajinkya, S. A. (2015). Prevalence of mobile phone dependence in secondary school adolescents. *Journal of Clinical and Diagnostic Research, 9*(11), 6-9.
11. Toda, M., Monden, K., Kubo, K., & Morimoto, K. (2006). Mobile phone dependence and health-related lifestyle of university students. *Social Behavior and Personality, 34*(10), 1277-1284.
12. Ministry of Science and ICT & National Information Society Agency (2016). *The survey on smart phone overdependence*. Seoul: Ministry of Science and ICT & National Information Society Agency.
13. Billieux, J., Van der Linden, M., D'Acremont, M., Ceschi, G., & Zermatten, A. (2007). Does impulsivity relate to perceived dependence on and actual use of the mobile phone? *Applied Cognitive Psychology, 21*, 527-537.
14. Lopez-Fernandez, O., Honrubia-Serrano, L., Freixa-Blanzart, M., & Gibson, W. (2013). Prevalence of problematic mobile phone use in British adolescents. *Cyberpsychology, Behavior, and Social Networking, 17*(2), 91-98.
15. Shih, D., Chen, C., & Chiang, H. (2009). An empirical studies on mobile phone dependency syndrome. In Proceedings of the 8th International Conference on Mobile Business, 176-181.
16. National Youth Policy Institute (2010). *The 2010 Korean children and youth panel survey I project report*. Seoul: Korea National Youth Policy Institute.
17. Fuster-Parra, P., Tauler, P., Bennasar-Veny, M., Ligeza, A., Lopez-Gonzalez, A. A.& Aguilo, A. (2016). Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Computer Method and Programs in Biomedicine, 126*, 128-142.
18. Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software, 35*, 1-22.
19. Norsys Software Corporaton (2018). Netica is a trademarks of Norsys software Corporation, https://www.norsys.com/netica.html, last accessed 2018/02/12.

Learning the required entrepreneurial best practices using data mining algorithms

Waseem Ahmad[1,2], Shuaib K. Memon[3], Kashif Nisar[4], Gurpreet Singh[1]

[1] Toi Ohomai Institute of Technology, New Zealand, [2] AGI Education Limited, New Zealand, [3] Auckland Institute of Studies, New Zealand, [4] Knowledge Technology Research Unit, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Malaysia
{waseem.ahmad, gs2042}@toiohomai.ac.nz, shuaibm@ais.ac.nz, kashif@ums.edu.my

Abstract. In this research, our focus is to establish a relationship between some of the entrepreneurial best practices such as good networking skills, developing a clear vision, perseverance and ability to take risks with the business success in the field of kiwifruit contractors. Failures at the initial stage of this business is a common occurrence in the Bay of Plenty region of New Zealand. For aspiring kiwifruit contractors achieving success is a herculean but a possible task. The success factor in this research is calculated based on the number of hectares cultivated land and the number of employees hired by the contractors. The research design adopted in this study is the quantitative research approach, the instrument of a well-structured questionnaire was devised, which was based on the 5 point Likert scale format. Weka, a well known data mining toolbox was used for the analysis of primary data collected from the respondents. In this research, rule based and decision tree algorithms were used to extract useful and actionable information from the data. The study concluded that clear vision and risk taking capabilities are two most important features required to become successful in this business.

Keywords: Entrepreneurial skills, Data mining, Networking skills, Rule based Learning Algorithms, Decision Tree learning

1 Introduction

Small and Medium Enterprises (SMEs) are considered as a major growth engine behind any economy. According to D'Imperio [1], approximately, 95% of the businesses in the world are SME's, which employ almost 60% workforce in the private sector. SME's contribute at least 16% to the GDP of low income countries, however on the other hand they contribute about 51% to the GDP of high income countries. In the case of New Zealand 97% of the enterprises are SME's, 60% of the SME's die before five years of their establishment [2].

Bay of Plenty region of the New Zealand has always been associated with the production of kiwifruits. In a survey conducted by "Te-Ara" in 2012, Bay of Plenty re-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_45

gion proudly had a monopoly on the total cultivated hectares of kiwifruit in New Zealand. Approximately 77.7% of the kiwifruit hectares were cultivated in the Bay of Plenty region (9,912 hectares of New Zealand's total of 12,757). This achievement is a result of the efforts made by the various SME's i.e. Kiwifruit contractors.

For a successful and ongoing operation of this lucrative horticulture practices, a huge amount of skilled labour force is required, hence, kiwifruit contractors (SME's) comes into the picture. They are the major suppliers of labour force in the kiwifruit orchards in the Bay of Plenty region. This kiwifruit contractors business is mainly dominated by the migrant population of New Zealand. Therefore, it is important to investigate some of the entrepreneurial best practices in successfully starting a kiwifruit Contractor Business in the Bay of Plenty Region of New Zealand. Over a period of time, various kiwifruit contractors have emerged; many have made fortunes in this lucrative business. These veterans are well known for their lavish life-styles in the Bay of Plenty region. Attracted by their lifestyle, a new army of kiwifruit contractors known as fresher kiwifruit contractors; are trying to establish their foothold in this lucrative domain. The objective of this research is to assist these fresher kiwifruit contractors to find essential entrepreneurial skills required to become successful in this business.

Since, the cultivated land mass (kiwifruit hectares) remains the same, it has become very difficult for new/fresher kiwifruit contractors to attract new business, because the veteran kiwifruit contractors still have the monopoly in this domain. There is a very small number of kiwifruit orchard owners and these owners have been working with veteran contactors for decades. Therefore, it is a real challenge for a new contactor to get into this business. Ideally, a company life cycle is divided into seven stages: (1) seed, (2) start-up, (3) growth, (4) established, (5) expansion, (6) maturity and (7) exit/rebirth or death in some cases [3]. Here the scenario is very much the paradox to the ideal situation i.e. the SME's like kiwifruit contractors are more vulnerable to die out at a very early stage of the company life cycle. It has been well established that some essential traits such as *"better networking skills, ability to take calculated risks, perseverance, and ability to convince the stakeholders, developing vision, and personal recommendations"* can assist new kiwifruit contactors to get establish in this business [15, 16, 17].

This research will help the fresher kiwifruit contractors to get a fair chance in this lucrative domain to establish their presence among the kiwifruit orchard owners. This research tries to uncover the relationship between various entrepreneurial best practices (such as networking skills, perseverance, developing a clear vision and ability to take risks) with the level of success achieved in the kiwifruit contractors business.

2 Literature Review

SMEs' contribute almost 97.2% of the total business in New Zealand; revenue generated from SME's contributes 42% to the total GDP of the nation. According to Fox et al. [2] SMEs' provide jobs to about 30.5% of the total workforce. Kiwifruit contractors fall under SME's. Very little or no research has been conducted on this topic.

Only by the help of comparative studies, this research can be advocated further. Comparative study is a tool through which we can correlate the previous findings of various authors with the research area we are focusing on.

Stokes & Lomax [4] conducted a survey on different small and medium businesses, for the role of word of mouth in acquisition of customers. It was during this research, influence of the entrepreneur on the overall wellbeing of the business, which leads to positive word of mouth, among the target audience was closely observed, by the instrument of various face to face interviews and questionnaire. In other words type of leadership and right set of strategies decides the future of any business. The result of this research confirmed the importance of word of mouth in acquiring customers. Two aspects of word of mouth also came into picture i.e. input word of mouth which relies upon sources & types and output word of mouth which rely upon target audience & content. This research has proved that, positive word of mouth can act as an effective tool for developing favourable marketing strategies for SME's. Similarly, a thorough research will be conducted on the kiwifruit orchard owners, to establish a relationship between the role of positive word of mouth and the outsourcing of horticulture activities to kiwifruit contractors.

Neti [5] mentioned that social media is a platform by virtue of which, companies (small or medium) get a chance to do marketing of their products and services without the involvement of any middleman. Making use of this technology connects the customers directly with the organization. In other words, it is on-line word of mouth. It is good marketing platform wherein companies can promote themselves by the instrument of advertisement over World Wide Web. Similarly, the fresher kiwifruit contractors can also go online for the purpose of promotions. As all the kiwifruit orchard owners are well connected over World Wide Web. It has become, easy to approach orchard owners over internet for getting business. Under this research an experiment establishing the relationship between acquiring customers through advertisements done through social media will be conducted. The results obtained from this experiment will be analysed closely and shared with budding kiwifruit contractors for their future prospects. A "two step flow model" developed by Katz & Lazarsfeld [6] can be taken for reference to reinforce this research with proof.

Goldsmith [7] concluded that social communication, opinion leadership and word of mouth influence the consumer behaviour. This is a phenomenon by virtue of which communication that occurs between the consumers decides the goodwill of a business. The need for off-line social communication has always been associated with the customer acquisition or customer turnover [8]. Moreover, sometimes people are influenced by the thinking of an opinion leader, similarly, an interview can be conducted on the representative of the NZ kiwifruit association in Bay of Plenty. By conducting this interview, we can establish the relationship between leadership and it influence on individual decision making of kiwifruit orchard owners for outsourcing of their horticulture activities. Every business has unique nature and requirements; hence, there is a need for custom made marketing strategies for SMEs. Similarly, Kiwifruit contractors business demands for the identification of these unique requirements, based on which appropriate marketing strategies can be formed. In order to identify these unique requirements, a face to face interview will be conducted with fresher and veteran kiwifruit contractors and kiwifruit orchard owners. Following questions will be answered through this research:

- Does developing a clear vision can help the fresher contractors to establish themselves in this lucrative domain?
- To what degree entrepreneurial skills such as ability to take risks and perseverance are helpful in becoming successful kiwifruit contractor?
- Does experience in the kiwifruit industry is helpful in achieving success in this domain?

3 Research Methodology

For the successful conduct of this research the instrument of questionnaire was used. The success in the kiwifruit contractors business was kept as the dependent variable which was measured by following two factors:

1. Total cultivated land area in hectares by each contractor in the current year.
2. Total number of employees hired by each contractor in the current year.

The independent variables were divided into six sections, four out of six categories were related to entrepreneurial best practices. The independent variables used in this study are as follows:

3. Total experience in kiwifruit industry
4. Number of years in kiwifruit contractor business.
5. Role of networking skills
6. Role of developing a clear vision
7. Role of Perseverance
8. Role of risk taking ability

According to Roscoes [9] rule of thumb is that any data gathered between 30 and 500 are good for any survey, moreover it also states that 10% size of the parent population can represent the entire population. There were approximately 230 registered kiwifruit contractors in the Bay of Plenty region, out of these 40 kiwifruit contractors responded to this questionnaire. These 40 respondents varied in age groups, educational backgrounds, country of origin and experience in kiwifruit business. The data was gathered using randomly sampling method. A number of field tours were conducted to gather the data required for this research. The description of the questions asked in this survey is highlighted in Table 1. A 5-point Likert scale questionnaire was developed based on the points stated in Table 1. The average score of each category (networking skills, vision, Perseverance and risk taking) was used to describe the score achieved in that category.

In this paper three data mining algorithms, namely, OneR, Modlem and J48 are used to extract useful information from the data. OneR algorithm [10] generates single level decision tree. The algorithm computes confusion matrix against all the variables and selects a single variable that has highest classification accuracy. In the literature researchers have used this algorithm as a baseline model due to its simplicity and ability to produce reasonably accurate results. J48 is a decision tree algorithm proposed by *Quinlin* [11]. This algorithm uses the concept of entropy or information gain to construct a decision tree from the dataset. At each node, algorithm selects a variable that has highest information gain to expand the decision tree. The third algorithm used in this paper is Modlem. This algorithm generates rules based on rough set theory and more information on this algorithm can be found in [12].

Table 1. Description of the questionnaire used in this research

Criteria	Questions
Networking Skills	• Social media plays an important role in getting business in the New Zealand kiwifruit industry. • Print media such as news-papers and distribution of fliers helps in getting new business. • Regular follow ups are necessary to get new business. • Adopting a formal approach to deal with potential clients helps in getting business. • Adopting an informal approach to deal with potential clients helps in getting business.
Developing a Clear Vision	• Developing clear short and long term goals help in getting new contracts. • By communicating our future goals to the staff members, increase the likelihood of getting new contracts. • Core values should be aligned directly with the future goals. • Core values increase the chances of acquiring new contracts.
Perseverance	• New business contracts encourage us to focus more on my business. • Failure in acquiring new contracts affects the level of commitment towards business. • When required I find new ways to overcome the failures.
Ability to Take Risks	• Compared to other kiwifruit contractors, I take more financial risks. • Compared to other kiwifruit contractors, I take lower financial risks. • I would be willing to risk a significant percentage of my income in order to get good return on investment. • I would accept potential short term losses in order to pursue my long term goals. • Taking calculated financial risk is important for my business. • According to me higher the risk meaning higher the return on the investment.
Success	• Total cultivated area (in hectares). • Total Number of Employees.

4 Experimental Results

The collected data has 7 features including class labels. Expectation Maximisation (EM) clustering algorithm [13] was used to find naturally occurring clusters in the data. EM clustering algorithm produced 2 clusters in the data. Therefore, numeric success data was converted into two class nominal data. These two classes (0 and 1) can be regarded as 'less successful contractors' and 'highly successful contractors' (Table 2). There were overall 40 data samples, out of which 27 were belonging to less successful contractors and 13 instances to highly successful contractors' class. The class variable was obtained by considering cultivated area and number of employees (eq. 1).

$$class\ variable\ (numeric) = \log\ (cultivated\ area * Number\ of\ employees) \qquad (1)$$

Table 2. Numeric success class variable transformed to nominal labels

Success variable range	Class labels
0.00 - 3.70	0
3.71 - onward (maximum value was 4.78)	1

All experiments are performed using Weks (a data mining tool) [14]. In Table 3, model accuracies of three algorithms (OneR, Modlem and J48) are compared using full and partial datasets. The models show accuracies on the training data and using 10-folds cross validation. The cross validation approach is used to demonstrate the robustness of the acquired models on the unseen data. Three variations (scenarios) of the dataset were used to evaluate the performance of algorithms.

- **Scenario 1:** Full dataset (all six variables) were used to build models.
- **Scenario 2:** Kiwifruit experience attribute is removed from the full data
- **Scenario 3:** Kiwifruit experience and contractor experience attributes are removed from the full data.

The Table 3 indicates that Modlem algorithm has outperformed other two algorithms. However, all three algorithms have comparable 10-folds cross validation results (around 75%) at scenario 3.

Table 3. model accuracies achieved on OneR, Modlem and J48 algorithms using training and 10-Folds cross validation results.

Algorithms	OneR		Modlem		J48	
Accuracies (%)	Training	10-Folds	Training	10-Folds	Training	10-Folds
Scenario 1	87.5	77.5	100	82.5	92.5	82.5
Scenario 2	87.5	77.5	100	80	92.5	75
Scenario 3	85	75	97.5	77.5	92.5	75

The models built using OneR algorithm can be seen in Table 4. The model built on full dataset has an accuracy of 87.5% whereas, model built by excluding kiwifruit experience had an accuracy of 85%. On the full dataset, kiwifruit experience attribute came out the most important variable and approximately 88% of the information in the data can be explained by this single variable. On the other hand, if partial dataset was used (excluding experience variables), then vision is the most stand out variable for the success of the business (85% of the data can be explained using vision variable alone). This finding indicate that to become successful in this business fresh contractors either need more than 20 years of experience or in the absence of the experience, they must have very strong business vision.

Table 4. Rules obtained using OneR algorithms

Full Dataset (Training accuracy 87.5%)	**Scenario 3 (Training accuracy 85%)**
Kiwifruit Experience < 18.5 → 0 (class label)	Vision < 4.625 → 0 (class label)
Kiwifruit Experience < 37.5 → 1 (class label)	Vision >= 4.625 → 1 (class label)
Kiwifruit Experience >= 37.5 → 0 (class label)	

Second algorithm used in this paper is Modlem. This algorithm produced an accuracy of 100% and 97.5% on full and partial datasets respectively. The rule generated using both datasets can be seen in Tables 5 and 6. The rules indices 1, 3 and 5 in Table 5 are consistent with the results obtained using OneR algorithm in Table 4. Rule 4 of the Table 5 shows a very interesting trend. According to this, a contractor must have very good networking skills otherwise, even more than 20 years of the experience in the kiwifruit sector will still fall him in less successful contactor category.

There are some interesting findings in the Table 6, where partial data was used to construct rules. According to rule 5, high scores in perseverance and networking skills do not qualify for success in the kiwifruit contractor business. Rule 6 (Table 6) is consistent with the findings in Table 4, where high score in vision leads to success in the business. Lastly, rule 8 of Table 6 suggests that a contractor must have high risk taking capabilities along with moderate level of networking skills to become successful in this business.

Table 5. Modlem rules extracted on full dataset (training model accuracy 100%)

Rule 1. (KiwifruitExperience < 17.5) & (Vision >= 3.63) => (class = 0) (17/17, 62.96%)
Rule 2. (RiskTaking < 2.75) => (class = 0) (9/9, 33.33%)
Rule 3. (KiwifruitExperience < 11.5) => (class = 0) (9/9, 33.33%)
Rule 4. (NetworkingSkills < 3.7) & (KiwifruitExperience >= 21.5) => (class = 0) (2/2, 7.41%)
Rule 5. (Vision >= 4.63) => (class = 1) (7/7, 53.85%)
Rule 6. (KiwifruitExperience >= 18.5) & (Perseverance >= 4.17) => (class = 1) (2/2, 15.38%)
Rule 7. (KiwifruitExperience in [17.5, 21.5]) & (Vision >= 4.38) => (class = 1) (4/4, 30.77%)
Rule 8. (Vision < 3.63) & (ContractorExperience >= 6) => (class = 1) (1/1, 7.69%)

Table 6. Modlem rules extracted on partial dataset (scenario 3) with training model accuracy of 97.5%

Rule 1. (RiskTaking < 2.75) => (class = 0) (9/9, 34.62%)

Rule 2. (NetworkingSkills < 3.5) => (class = 0) (5/5, 19.23%)

Rule 3. (Vision < 4.38) & (NetworkingSkills < 3.9) => (class = 0) (13/13, 50%)

Rule 4. (Perseverance < 3.5) => (class = 0) (2/2, 7.69%)

Rule 5. (Perseverance >= 4.17) & (NetworkingSkills >= 4.1) => (class = 0) (8/8, 30.77%)

Rule 6. (Vision >= 4.63) => (class = 1) (7/7, 58.33%)

Rule 7. (Vision >= 4.38) & (RiskTaking < 3.75) => (class = 1) (3/3, 25%)

Rule 8. (RiskTaking >= 4.08) & (NetworkingSkills in [3.7, 4.1]) => (class = 1) (3/3, 25%)

Rule 9. (Vision in [3.38, 3.63]) & (NetworkingSkills >= 3.9) => (class = 1) (1/1, 8.33%)

Rule 10. (RiskTaking in [2.92, 3.08]) & (NetworkingSkills >= 3.5) => (class = 1) (1/1, 8.33%)

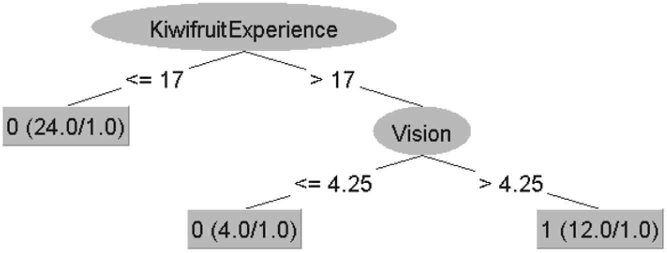


Fig. 1. Decision tree generated by J48 algorithm using full dataset with training data accuracy of 92.5%

According to Figure 1:

- **If** Kiwifruit experience is less or equal to 17, then the class label is '0'
- **Else If** Kiwifruit experience is greater than 17 and Vision is less or equal to 4.25, then the class label is '0'
- **Else If** Kiwifruit experience is greater than 17 and Vision is greater than 4.25, then the class label is '1'

According to Figure 2:

- **If** Vision is less or equal to 4.5, then class label is '0'
- **If** Vision is greater than 4.5, then class label is '1'

These two rules (in Figure 2) cover 80% of the data and remaining 20% of the data can be classified using vision and risk tasking variables. The rules obtained using J48 are consistent with the rules extracted using OneR and Model algorithms.

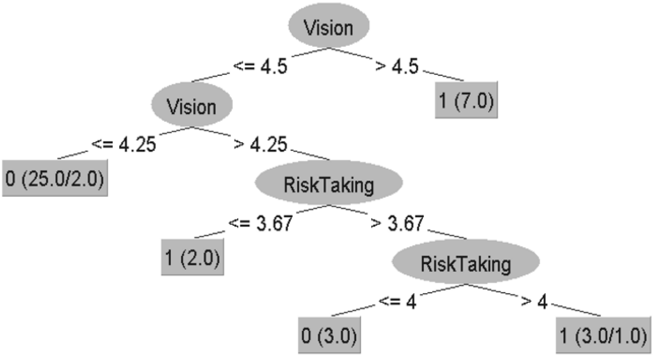


Fig. 2. Decision tree generated by J48 algorithm using partial dataset (scenario 3) with training data accuracy of 92.5%

5 Conclusion

This this paper, we have looked at the various entrepreneurial skills variables and experience in the kiwifruit industry to measure the success of kiwifruit contractors. Our analysis demonstrate that years of experience directly translate into the success in the contractors business. In our experiments, all three algorithms suggested that approximately more than 18 years of experience is required in this business to become successful contractor. This finding suggests that kiwifruit contractors must spend many years to earn orchard owners respect and subsequently their business. According to J48 algorithm, kiwifruit contractor must have more than 17 years of kiwifruit experience with vision score of more than 4.25 to become successful contractor in this field. Moreover, when experience was removed to construct the decision tree (scenario 3), vision and risk taking variables were adequate to build the model. This finding suggests that to become successful in this business, new contractors must have strong business vision along with high risk taking capabilities. These research outcomes can help aspiring kiwifruit contractors to use right set of entrepreneurial skills to become successful in kiwifruit business.

When it comes to outsourcing of horticulture activities the orchard owners most of times go for veteran contractors to guarantee good yield of crops. This mind set of kiwifruit orchard owners needs a makeover to save fresher kiwifruit contractors. Whenever there is a huge amount of money at stake, no-one dares to take risk with a fresher; however, in case of budding kiwifruit contractors this risk can be minimized.

References

1. D'Imperio, R. (2012). Growing the global economy through SMEs. Retrieved April 10, 2016, from http://www.edinburgh-group.org/media/2776/edinburgh_group_research_-_growing_the_global_economy_through_smes.pdf

2. Fox, A., Sun, P., & Stewart, I. (2012, November 5). Where is NZ Inc. going wrong? Retrieved September 14, 2015, from http://www.stuff.co.nz/business/small-business/7906892/Where-is-New-Zealand-Inc-going-wrong
3. Janssen, T. (2014). The 7 stages of business life cycle. Retrieved April 27, 2016, from http://www.justintimemanagement.com/en/The-7-stages-of-business-life-cycle
4. Stokes, D. and Lomax, W. (2002). Taking control of word of mouth marketing: the case of an entrepreneurial hotelier. Journal of Small Business and Enterprise Development. 9(4), 349-357
5. Neti, S. (2011). Social Media and its Role in Marketing. Retrieved October 2, 2015, from http://www.ijecbs.com/July2011/13.pdf
6. Katz, E., & Lazarsfeld, P. (1955), Personal Influence, New York: The Free Press.
7. Goldsmith., R. (2008). Electronic Word-of-Mouth. Retrieved September 16, 2015, from http://www.igi-global.com/chapter/electronic-commerce-concepts-methodologies-tools/9610
8. Lazarsfeld, P.F., Berelson, B. & Gaudet, H. (1944). The people’s choice: How the voter makes up his mind in a presidential campaign. New York: Columbia University Press.
9. Roscoe, J.T. (1975) Fundamental Research Statistics for the Behavioural Sciences, 2nd edition. New York: Holt Rinehart & Winston.
10. C.R., H., Very simple classification rules perform well on most commonly used datasets. Machine Learning, 1993. **11**: p. 63-91.
11. Quinlin, J.R., *C4. 5: programs for machine learning.* Morgan kaufmann, 1993.
12. Stefanowski, J., *On rough set based approaches to induction of decision rules.* Rough sets in knowledge discovery, 2008. **1**(1): p. 500-529.
13. Fraley, C. and A. Raftery, How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Computer Journal, 1998. 41(8): p. 578-588.
14. WEKA, Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java.
15. Namrata, C., and Niladri D. (2016). A Study on the Impact of Key Entrepreneurial Skills on Business Success of Indian Micro-entrepreneurs: A Case of Jharkhand Region . Global Business Review , 17, 1, 226-237
16. Dimov, D. (2007). Beyond the Single-Person, Single-Insight Attribution in Understanding Entrepreneurial Opportunites. Entrepreneurship Theory and Practice. 713-731.
17. Van Gelderen, M. (2012). Perseverance Strategies for Enterprising Individuals. International Journal of Entrepreneurial Behaviour & Research. Emerald Publishing Group. 1-35.

Agent Based Irrigation Management for Mixed-Cropping Farms

Kitti Chiewchan, Patricia Anthony and Sandhya Samarasinghe

Lincoln University, Christchurch 7608, New Zealand
Kitti.Chiewchan@lincolnuni.ac.nz
Patricia.Anthony@lincoln.ac.nz
Sandhya.Samarasinghe@lincoln.ac.nz

Abstract. This paper describes the development of an intelligent irrigation management system that can be used by farmers to manage water allocation in the farms. Each farm is represented as a single agent that can work out the actual water required for each crop in the farm based on the crop's drought sensitivity, growth stage, the crop coefficient value and the soil type. During water scarcity, this system can prioritise irrigation allocation to different crops on a farm. Our initial experiment showed that using the irrigation management system, the farm can achieve a consistent water reduction which is more than the required reduction. The results showed that the agent consistently recorded water reduction higher than the actual reduction required by the water authority. This significant reduction means that more water can be conserved in the farm and reallocated for other purposes.

Keywords: agent-based model, water allocation, utility function, water reduction.

1 Introduction

Water use and water demand have increased steadily in New Zealand over the last 20 years resulting in insufficient water availability. The water usage data [13] shows that Canterbury water allocation makes up 58% of the New Zealand's total water allocation where it contributes 70% of the New Zealand irrigated land. It is expected that water demand will become a problem in the future because the irrigated areas in Canterbury have been increasing for the last 13 years (from 300,000 ha in 2002 to 500,000 ha in 2015) [7]. This demand directly affects the water allocation scheme in Canterbury. Currently, the water usage policy is based on "first in, first served", which means request for water consents are processed and determined in the order they are received. This policy worked in the past because water capacity and farming areas are in equilibrium. Unfortunately, "first in, first served" system is not the most efficient way to manage water. This is due to the fact that even though water demand has increased over the years, water capacity remains unchanged [13]. As the irrigation is based on estimate, there is a possibility that crops received more water than necessary leading to wastage. Water needs become more serious during drought season and

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_46

so it is very important to conserve water and prioritise crop water need such that high yield crops get the highest priority so as not to affect productivity. If farmers can decide on the irrigation plan that is dependent on the importance of the crops, they can reduce water need in the farm and reduce the loss in productivity during water restriction [2].

To assist in the irrigation planning, farmers often used computing tools and two of the most common ones are OVERSEER and IrriCalc. IrriCalc is an irrigation management tool for irrigation water requirement. It can determine the irrigation water need due to seasonal planning [6]. OVERSEER is owned and supported by the Ministry for Primary Industries. OVERSEER's model uses daily soil water content data to calculate the daily water drainage. IrriCalc and OVERSEER require input such as selected month, farm location, and type of irrigation system for daily water need calculation. However, IrriCalc and OVERSEER use different models to calculate climate data. OVERSEER assumes full 'canopy' cover (value of 1) whereas IrriCalc uses a seasonally adjustable value, with an average value of 0.8 [16]. There are other computing tools such as APSIM (Agricultural Production Systems Simulator) and AquaTRAC. APSIM is a modeling framework that contains a suite of modules to enable simulation of agricultural systems. It provides a set of modules (physical process in farm, farm management rule, simulation engine) to support higher-order goal of farming simulation. It provides accurate predictions of crop production based on climate, genotype, soil and management factors. On the other hand, AquaTRAC is a software program developed by Foundation for Arable Research (FAR) which assists cropping farmers with their irrigation scheduling. It calculates when and how much irrigation to apply to optimise yield for each crop by including data on crop type, soil type, weather and irrigation levels [10][16]. However, these tools are limited to calculating water requirement for a single crop in a farm and is unable to address water requirement for farm with multiple crops. Hence, there is a need for a better irrigation management that can accurately estimate and manage irrigation water on the farm either for single crop farms or mixed-crop farms. This paper proposes an agent-based irrigation management system that can be used to allocate water efficiently in the farm based on the farm's characteristics. The remainder of the paper is organised as follows. Section 2 describes the irrigation management and its application in New Zealand, the crop water needs calculation and related works on agent-based irrigation management. The proposed agent-based model for intelligent irrigation management system is discussed in Section 3. We present the experiment and result in Section 4 and finally Section 5 concludes and discusses future works.

2 Related Works

2.1 Irrigation scheme and the process

There are three stages in the cycle of crop growth; 1) soil preparation 2) irrigation process and 3) after irrigation process. During the soil preparation, farmers need to decide the location of the irrigation, the water capacity and the irrigation schedule to prepare for planting [15]. During the irrigation process, farmers need to check and work out their irrigation plan for the whole agriculture areas by making references to the weather, season and water policy. The after irrigation process cycle focuses on improving soil quality after the irrigation season and improving irrigation for the next seasons. The Evapotranspiration Rate (ET) is an important variable in irrigation which relates to land location, soil type, and planting season in the farm. ET is the summation of evaporation and plants transpiration from soil to atmosphere. To ensure that each crop gains the highest yield, maximum water need must be applied. This irrigation water need can be estimated using ET and another variable called the crop coefficient (Kc). The value of Kc is determined based on the crop growth stages which are initial state, crop development stage, mid-season stage and late season stage. The water need for each crop varies from one crop stage to another.

2.2 Calculating Crop Water Need

The irrigation water need is defined by evapotranspiration (ET) which is the water transformation process from land to atmosphere by evaporation. Crop water need can be calculated using the following formula [9]:

$$ET_{crop} = ET_0 \times K_C \qquad (1)$$

Where: ET_{crop} = Crop water need
ET_0 = Influence of climate on crop water need
K_C = Influence of crop type on crop water need

The common tools used in irrigation management (such as OVERSEER, IrriCalc, APSIM, AquaTRAC) follow this formula to estimate the crop water need in the planting season. However, these tools do not consider the drought sensitivity of different crops and drought sensitivity based on crop growth stages. Drought sensitivity is a crop characteristic under drought stress where they need more water for every growth stage to ensure maximum productivity such as paddy rice and potato. If various crops are grown on an irrigation scheme, it is advisable to ensure that the most drought sensitive crops get the highest priority.

2.3 An agent-based approach to irrigation

Agent-based Programming (AP) is a software paradigm that uses concepts from Artificial Intelligence (AI). Agent's behavior depends on what it is tasked to do and gathers information about the environment to make a decision. AP has been used to solve resource allocation problems [17]. A software agent is essentially a special software component that can operate without the direct intervention of a human. These agents when grouped together, form a multi-agent system that can be used to model and solve complex systems as it has the ability to introduce conflicting goals and act upon it. An agent senses and reacts to the changes in the environment. An agent is able to exhibit goal-directed behavior by taking initiative while ensuring that its goal is achieved. The agent can learn and adapt itself to the demands of its users and fit its environment [5].

Agent-based programming has been used in water resource allocation. For example, [8] used agent-based modeling to simulate the interaction between farmers who are stakeholders in transboundary Nile River. This simulation generated farmer agent and water sharing scheme from water usage behavior. This model was developed to optimize allocated water for each user with different water requirement to find a fair water allocation for stockholders at Nile river basin.

An agent-based model was applied to the history of irrigated agriculture investigation in Spain. The purpose of this study was to study the impact of farmers' characteristics on land-use change and their behavior of groundwater usage [12]. They showed that agent-based model can be utilized to enhance this understanding even when data is scarce and uncertain. An agent-based model was used to simulate irrigated system in Senegal River Valley to find the limitation of water used based on behavioral factors (resource capacity, a set of individual water used rule, and a set of collective rules) [3]. The focus of this work was to verify that MAS is a suitable architecture that can be used to theoretically study irrigated systems' viability. Using MAS, they designed and developed virtual irrigated systems as alternative to real labs.

A simulation based on multi-agent system was developed to study and analyse the collective action when a certain water policy is changed [4]. This model was able to capture the problems related to collective action in water markets in different scales infrastructure provision. The system was also used to simulate the behavior of different water users to represent social and institutional relations among users. This work demonstrated how MAS can be used to understand water use and the complexity of water use within sub-basins.

An agent-based model was developed to simulate different water users/ behaviors as well as their reactions to different conflict scenarios in water usage scheme. They simulated the behavior and interactions of the conflicting parties and modeled it as a game. This model was used to explain the interactions between parties and to enable decision making among the stakeholders [1].

Giuliani et al. [11] developed a multi-agent system to design a mechanism for water management. The agent-based model represents the interactions between the decision makers to demonstrate a hypothetical water allocation problem, involving several active human agents and passive ecological agents. They used different regulatory mechanisms in three different scenarios of water availability to investigate efficiency-acceptability tradeoff. The results obtained showed that this approach was able to support the design of the distributed solution.

Zhao et al. [18] compared the water user behavior under the administered system and market-based system by developing an agent-based modeling framework for water allocation analysis. Their analysis showed that the behaviors of water users were dependent on factors such as transactions, administrative cost and costs.

Overall, irrigation management is a complex problem because it is hard to determine the water need in the farm as there are many dynamic factors that need to be considered. For example, crops growth are in different stages, temperature changes on a daily basis, the soil moisture varies, and sometimes there is a prolong drought season. Moreover, different farms have different water requirements because of the varying crop type, soil type, farm location, and farm size. Agent-based programming has advantage over other software approaches because it can work with uncertain factors and supports non-linear data. Moreover, it is flexible and autonomous under complex situations.

3 Agent-based model for irrigation management

3.1 Conceptual design

This study focuses on using agent-based approach to optimise water allocation in mixed-cropping farms. The crop water need is calculated based on many factors such as crop water requirement, moisture on the ground, and farm types. It is assumed that an agent represents a single farm where each farm may have a single type of crop or mixed types of crop. The agent is able to estimate the crop water need on a daily basis based on the current state of the farm's characteristic (i.e. crop types, crop stage, soil type, etc.). The agent will also be able to work out the irrigation plan for the farm for a given planting season. During water scarcity (such as drought season), the agent will work out an irrigation plan that will prioritise crop water needs based on the prevailing condition.

The agent will be able to generate the irrigation plan on a daily basis, weekly or for the whole season. The crop information database contains an up-to-date information about the state of the farm including information about each crop, its growth stage, its location, size of the plot and soil type. During water scarcity, user may enter the percentage of water reduction required by the authority. The agent will then calculate the water need for each crop in the farm, calculate the expected utility for each crop, decide which crops have higher water need by prioritizing them and determine how much water should be reduced for each crop in the farm. This prioritization is determined based on crop's expected utility which takes into account the potential yield of the crop, the drought sensitivity and growth stage and the soil type. For example, if the farm contains high yield crops, more importance will be placed on the crop's yield to ensure that the revenue of the farm is not compromised. On the other hand, if there is no high yield crop on the farm, the other two factors can be considered.

Crops type can be divided into 2 categories: grazing pasture and other crops. High yield crops are crops that generate higher revenue for the farm. Examples of high yield crop include tomato, sugar cane and sugar beet. Most crops have four growing stages (germination, development, mid-season, late-season). Pasture has three stages (grazing, development and late-season). There are 3 types of drought sensitivity (low, medium, and high). Crop with high sensitivity to drought requires higher irrigation priority. Crops can be planted in plots with varying soil type (light, medium, heavy). Heavy soil can absorb more water and so it has low irrigation priority. On the other hand, light soil has bigger pores and absorbs less water. Therefore, light soil has higher priority over medium and heavy soil because of its inability to retain moisture in the soil moisture potential yield [14]. It is assumed that each type of crops is planted in plot of a certain size. This means that the water need for a crop is calculated to cover the plot that has been planted with that particular crop.

```
begin
   get water reduction percentage
   retrieve data information from crop characteristics database
   calculate the water requirement for each crop
   calculate total water requirement on farm
   calculate expected utility for each crop
   prioritise crops based on expected utility
   calculate water reduction for each crop
   generate water reduction plan for a farm
end
```

Fig. 1. The pseudocode of the irrigation management agent

The pseudo code for the water reduction calculation is shown in Fig.1. First, the user keys in the percentage of water reduction required to the system. The agent retrieves the crop information from the database, K_c value for each crop stage, and reference ET_0 data. Equation 1 is used to calculate the actual crop water requirement. The ET_0 value is retrieved from the New Zealand weather data and the K_c value is based on the FAO data reference. Agent will work out the individual crop water need and the total water requirement for the whole farm (this is the summation of individual crop water requirement). Next, the agent calculates the water reduction requirement for the farm based on the percentage of required reduction entered by the user. Then, the agent calculates the expected utility for each crop based on its properties. Once the crops are prioritized, the agent will estimate water reduction for each crop. Finally, agent generates irrigation reduction plan as the output to the farmer.

3.2 Prioritising irrigation water needs to multiple crops.

To calculate water requirement, agent uses ET_{crop} and plot size on each crop (C_{plot}) as follows:

$$C_{wrq} = ET_{crop} \times C_{plot} \qquad (2)$$

To prioritize irrigation water need to multiple crops, the agent used three determinants namely the drought sensitivity and the growth stage, the potential yield for crops and the soil type [2]. Each determinant is associated with a utility function that indicates the importance of that crop (the higher the value, the higher the irrigation priority). These utility functions (based on crop's potential yield, drought sensitivity and growth stage and soil type) are defined as follows:

$$f_{PY} = (C_{plot} \times C_{yield}) \times C_{price} \qquad (3)$$

$$f_{DS} = C_{stage} \times D_{val} \qquad (4)$$

$$f_{ST} = S_{type} \times C_{plot} \qquad (5)$$

Where:

f_{PY} = crop's potential yield function
f_{DS} = crop's drought sensitivity function
f_{ST} = soil type function
C_{yield} = yield amount of crop area
C_{price} = price per kilogram
C_{stage} = crop stage
C_{plot} = plot size of crop
D_{val} = drought sensitivity value
S_{type} = soil type

To determine the expected utility (EU) of each crop, the agent combines the three utility functions by allocating weights to denote their relative importance. Thus, the expected utility of each crop is calculated as follows:

$$EU_{crop} = \sum_{j \epsilon c} w_j f_j \qquad ; \ 0 \leq w_j \leq 1 \ and \ \sum_{j \epsilon c} w_j = 1 \qquad (6)$$

Where C is the set of determinant when agent works out the crop priority value, f is the utility function for each determinant and j is the individual parameter. At any time, the three utility functions can be considered by the agent for irrigation priority depending on what it sees as being important at that point in time.

For example, in an all pastures farm, the crop's drought sensitivity and growth stage determinant are the most important because the potential yield of the crop is the same (pasture only). On the other hand, if the farm is a mixed crops farm that has high yields crops, then the crop's potential yield is more important because is affect the productivity and farming income. Thus, a higher weight will be applied to the crop's potential yield function.

4 Experimental Evaluation

4.1 Experimental setup

To test the performance of our intelligent irrigation management system, we conducted three experiments using three farm setups, grazing pasture only farm, other crops only farm and mixed crops farm (which consists of pasture and other crops). These setups are common in New Zealand's farms. For each experiment, we randomly generate 100 farms with varying crop properties. For the pasture only farm, we set the crop type to pasture and randomise the growth stage to three different stages. In the other crops farm, we randomise the crop type and the growth stage to four different stages. The setup for the mixed crop farm is also similar to the other crop farm. However, we included pasture as the additional crop. The farm size is fixed to 200 hectare and the water capacity to 15,000 m^3. To validate the accuracy of our crop water need, we manually calculated the crop water need (the actual water need is calculated based on FAO's formula for calculating crop water requirement [9]) and compared this value with the value generated by our irrigation management system.

We run this experiment using four water reduction schemes at 5%, 10%, 15% and 20% and calculated the average difference between the actual reduction and the proposed reduction. This is to align with the real world setting where the water authority defines the water reduction percentage during water scarcity. We did not test it with reduction scheme higher than 20% as the water reduction scheme is usually capped at 20%. In the pasture only farm, we set the weight for the three determinants as (f_{PY}=0.2, $f_{ST} = 0.5, f_{DS} = 0.3$) to indicate the importance of growth stage and drought sensitivity. In the other crops farm, we set the weight as (f_{PY}=0.5, $f_{ST} = 0.3, f_{DS} = 0.2$) and in the mixed crops farms the weights were set to (f_{PY}=0.5, $f_{ST} = 0.3, f_{DS} = 0.2$).

5 Results and Analysis

The proposed water reduction by the system is shown in Table1 and Fig.2. In grazing pasture only farm, the average water reduction is much higher than the actual reduction for all cases. Our proposed irrigation management system recorded a reduction percentage of 9.36%, 11.79%, 16.70% and 20.81% for 5%, 10%, 15% and 20% reduction scheme. It can be seen that the agent was able to propose a higher water reduction than the actual reduction. In the other crops farm, the average water reduction is similar with all pasture farms for all cases (10.01%, 12.39%, 16.70% and 21.96% respectively).

Table 1. The average proposed water reduction for difference reduction schemes.

	Water reduction (m^3)							
	5% of reduction		**10% of reduction**		**15% of reduction**		**20% of reduction**	
	m^3	**%**	**m^3**	**%**	**m^3**	**%**	**m^3**	**%**
Actual water requirement	15,000	100	15,000	100	15,000	100	15,000	100
Actual reduction	750	5	1,500	10	2250	15	3000	20
Proposed reduction (pasture)	1,404.52	**9.36**	1,768.92.	**11.79**	2,504.5	**16.70**	3,121.46	**20.81**
Proposed reduction (multiple crops)	1,500.81	**10.01**	1,859.21	**12.39**	2,332.5	**16.70**	3,294.92	**21.96**
Proposed reduction (crops and pasture)	1,404.52	**9.36**	3,068.92	**20.46**	3,068.92	**20.46**	3,294.92	**21.96**

The result for the mixed crops farm also recorded a higher reduction percentage compared to the actual reduction. It recorded a reduction percentage of 9.36%, 20.46%, 20.46% and 21.96% for 5%, 10%, 15% and 20% reduction scheme. The results for the three types of farm are consistent across all the reduction schemes where all three recorded a higher than the actual reduction percentage. Based on this result, we can conclude that our proposed agent-based irrigation management system was able to consistently propose a significantly higher water reduction than the actual reduction required.

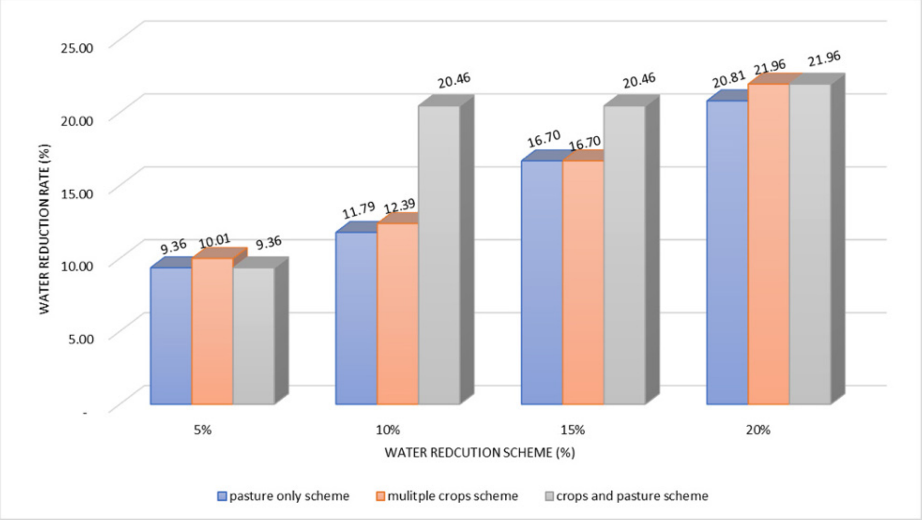


Fig. 2. The average water reduction based on water reduction scheme

6 Conclusion and discussion

In this paper, we describe an intelligent irrigation management system that makes water allocation decision based on the crop potential yield, the crop drought sensitivity and growth stage and the soil type. This tool is especially useful during water scarcity when farmers are required to make water reduction in the farm. Based on experimental result, it can be seen that the proposed model was able to save water even when water reduction is in place. It consistently proposed a water reduction plan that is higher than the actual reduction. For future work, we plan to extend this work by creating a community of agents that can work together to optimize water allocation in a community irrigation scheme. If each agent can accurately work out its crop water requirement in the farm, then it is quite possible that there is excess water that can be used for other purposes such as trading it with the other farmers in the community who might not have sufficient water. This will help the authority to maximize the allocation of water across the region. This water trading mechanism will also need to be further investigated.

References

1. Akhbari, M., & Grigg, N. S.: A framework for an agent-based model to manage water resources conflicts. Water resources management, 27(11), 4039-4052 (2013).
2. Anthony, P., & Birendra, K. C.: Improving irrigation water management using agent technology. New Zealand Journal of Agricultural Research, 1-15 (2017).
3. Barreteau, O., Bousquet, F., Millier, C., & Weber, J.: Suitability of Multi-Agent Simulations to study irrigated system viability: application to case studies in the Senegal River Valley. Agricultural Systems, 80(3), 255-275 (2004).
4. Berger, T., Birner, R., Mccarthy, N., DíAz, J., & Wittmer, H.: Capturing the complexity of water uses and water users within a multi-agent framework. Water Resources Management, 21(1), 129-148 (2007).
5. Bellifemine, F. L., Caire, G., & Greenwood, D.: Developing multi-agent systems with JADE. Vol. 7. John Wiley & Sons (2007).
6. Bright, J. C.: Prepared for Irrigation New Zealand. Aqualinc Research Limited, New Zealand (2009).
7. New Zealand statistics Homepage, https://www.dairynz.co.nz/, last accessed 2017/7/9.
8. Ding, N., Erfani, R., Mokhtar, H., & Erfani, T.: Agent Based Modelling for Water Resource Allocation in the Transboundary Nile River. Water 8(4), 139-151 (2016).
9. Doorenbos, Jan, Willian O. Pruitt, and A. Aboukhaled.: Crop water requirements. Food and Agriculture Organization, Rome, Italy (1997).
10. Foundation for Arable Research (FAR).: Irrigation management for cropping – a grower's guide, Australia (2010).
11. Giuliani, M., Castelletti, A., Amigoni, F., & Cai, X.: Multiagent systems and distributed constraint reasoning for regulatory mechanism design in water management. Journal of Water Resources Planning and Management 141(4), 04014068 (2014).
12. Holtz, G., & Pahl-Wostl, C.: An agent-based model of groundwater over-exploitation in the Upper Guadiana, Spain. Regional Environmental Change 12(1), 95-121 (2012).
13. Ministry for the Environment Homepage, https://www.mfe.govt.nz/sites/default/files/media/Fresh%20water/water-allocation-use-jun04.pdf last accessed 2018/3/20.
14. New Zealand Parliament Homepage, https://www.parliament.nz/resource/en-NZ/00PlibCIP151/431c33c3cf20b98103fa36e28a1dee1185801174 , last accessed 2018/3/9.
15. Williams, J. M., & Richardson, P.: Williams, J. Morgan, and Philippa Richardson. Growing for Good, Intensive Farming, Sustainability and New Zealand's Environment. Wellington, New Zealand (2004).
16. Wheeler, D.M. and Bright, J.: Comparison of OVERSEER and IrriCalc predicted irrigation and drainage depths. AgResearch. Report prepared for Overseer Management Services Limited, New Zealand (2015).
17. Wooldridge, M.: Agent-Based Computing. Interoperable Communication Networks 1, 71-97 (1997).
18. Zhao, J., Cai, X., & Wang, Z.: Comparing administered and market-based water allocation systems through a consistent agent-based modeling framework. Journal of environmental management, 123, 120-130 (2013).

A Review on Agent Communication Language

Gan Kim Soon[1], Chin Kim On[1], Patricia Anthony[2], Abdul Razak Hamdan[4]
[1]Center of Excellence in Semantic Agents, Faculty of Computing and Informatics, Jalan UMS, 88400, Universiti Malaysia Sabah, Sabah, Malaysia
[2]Faculty of Environment, Society and Design, Lincoln University, Christchurch, New Zealand
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, MALAYSIA
g_k_s967@yahoo.com, kimonchin@ums.edu.my,
patricia.anthony@lincoln.ac.nz, arh@ukm.edu.my

Abstract. Agent technology is a new emerging paradigm for software systems. In order to fully utilize the capability of this technology, multiple agents operate in software environment by cooperating, coordinating or negotiating with each other. However, these interactions require these agents to communicate with each other through a common language or protocol. Agent communication language (ACL) is a vital component in multiagent system (MAS) to enable the agents to communicate and exchange messages and knowledge. However, there are no universally agreed agent communication language that is widely adopted. Different agent communication languages and different semantic models have been developed to ease the communication between agents in MAS. The purpose of this paper is to review and highlight advances in the development of ACL.

Keywords: Agent Communication Language, KQML, FIPA-ACL, Mentalistic, Conversation Policy, Social Commitment

1 Introduction

Agent technology is a new emerging software paradigm that possesses certain characteristics that are suitable for computing environment which are highly heterogeneous, distributed and complex. To date, there is no universally agreed definition for what is an agent. There are many definitions used by different researchers to define agent in different research contexts. Genesereth [1] defined software agent as software component that is capable of exchanging knowledge and information. Bradshaw characterized software agent based on ascription and description. In ascription, software agent is defined based on its attribution and family resemblance whilst description software agent is defined based on a set of attribute list [2]. Nwana classified software agent into topology based on three primary intersect attributes which are cooperate, learn and autonomous [3]. Wooldridge defined agent as an autonomous software entity that is situated in some environment where it can monitor and response to changes proactively or reactively by itself or through communication with other agents to persistently achieve certain goal/task on behalf of user or other agents [4]. In [5],

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_47

agent definition is further distinguished between strong notion of agent and weak notion of agent. A weak agent is said to possess primary properties such as autonomy, reactive, proactive and social ability. Besides primary attributes, agent may also possess some of the secondary attributes such as benevolent, sincerity, rational, learnability and others. On the other hand, a stronger notion of agency is defined as possessing the mentality attitudes such as beliefs, desire, intentions and others. Hitherto, it can be observed that there is no single widely accepted definition of what is an agent. Nevertheless, there are some common properties that can be derived from these definitions such as agent is autonomous and it can communicate in order to exchange information. However, a single agent computation power is still limited to solve a large complex system which are decentralized and distributed.

In order to realize complex systems, the capability of a single agent is never enough. Thus, multiple homogeneous or heterogeneous agents are required to scale up for large, distributed complex systems. A system where multiple single agents work together is referred to as multiagent system (MAS). As mentioned MAS consists of multiple agents which may have common goal where they work together to achieve a certain task; or have self-interested goal where each agent competes against each other for resources; or they are required to coordinate to achieve a certain task. The advantages of MAS include scalability, efficiency, robustness and reusability [5]. However, in order to realize the interaction such as negotiation, cooperation, collaboration and coordination between agents, a common language protocol is required in order to achieve interoperability. These interactions can only be carried out if the agents can communicate and understand the communicated message syntactically and semantically.

Thus, agent communication language has been developed in order for agent to communicate with each other and understand the content of communication. In the next section, several existing agent communication languages will be discussed.

2 Agent Communication Language

Agent communication language (ACL) is a high level abstraction method for agent to exchange information and knowledge [1]. ACL allows more complex knowledge exchange such as plans, agent's goal, and believes which cannot be exchanged using object-oriented approach. Object oriented approaches such as remote procedure call (RMI), remote method invocation (RMI), COBRA and object broker request are not suitable for agent communication.

Knowledge manipulation and query language is the first ACL that was developed for agent communication [6][7]. It was initially developed as part of the knowledge sharing effort by DARPA with the aim to create a set of reusable tools to transfer and exchange high level knowledge and information [8][9]. Then it evolved to become high-level message oriented communication language between agents to exchange information and knowledge which is independent of the content syntax and ontology. KQML has three layers' organization structures that are composed of content, communication and message layer. The content layer represents the content of the message. The communication layer is composed of the message transport layer component such as sender and receiver. The message layer encodes the KQML message which includes wrapping the content and communication layer. The vital component of the message layer is the performatives. Performatives are utterance actions which are based on the speech act theory that denotes the illocutionary meaning of the speaker [10][11]. KQML's syntax is a LISP-like expression that is composed of performatives and pairs parameters and value. During the early development of KQML, there were no particular semantics models that were adopted and this has resulted in several variations of KQML dialect which were based on the application context. As a consequence, KQML has been criticised for its lack of formal semantic model which led to confusion and ambiguity in performatives meaning [12]. Although, KQML allow arbitrary content language, the de facto content language for KQML is the knowledge interchange format [13]. KIF uses the first-order predicate calculus to describe things in knowledge representation. The Ontolingua is used as the ontology for the KQML communication [14]. Labrou and Finnin deviced a semantic model for KQML based on precondition, post condition and complete condition of mental states [15][16][17]. However, this mentalistic notion suffers certain drawback which will be discussed in Section 4.

FIPA-ACL is an agent communication language specification developed by FIPA (Foundations of Physical Agents). FIPA is a non-profit organization formed by various organizations from academics to industry. The aim of FIPA is to develop a set of standards or specification to promote the interoperability of agent technology. To date, FIPA has produced a set of standard specification that needs to be adopted in order for agent to communicate and interact in interoperability mode. Among these set of specifications, one of the specification is called FIPA-ACL specification which promotes the FIPA-compliant agent communication language. The first FIPA-ACL specification was in 1997 and subsequently revised in 1998 and improved in the 2000 specification. [18][19]. The syntax of FIPA-ACL is similar to KQML syntax. The semantic model adopted is based on Cohen and Lévesque [20] which is an enhancement of Sadek's work in Arcol [21]. In FIPA-ACL, the performatives are known as communication acts. FIPA defined a set of communication acts in the FIPA communicative acts library specification which is based on the speech act theory [22]. FIPA-ACL does not constraints the use of new communication acts. However, in order to preserve interoperability, these communicative acts must be agreeable by communicating agents in both syntax and semantic. The semantic of FIPA-ACL is based on the feasibility precondition and rational effect. FIPA-ACL does not limit the content language that can be used but FIPA-SL has become the de facto standard for the FIPA-ACL [22]. FIPA-SL is based on the quantified multimodal logic made up of the belief, desires, uncertain beliefs and intentions modal operators. Other supportive specifications for the FIPA-ACL include FIPA ontology specification and FIPA interaction protocol specification which can be found in FIPA website [23]. FIPA-ACL also suffered several drawbacks such as no standard parser or reasoner for FIPA-SL and was criticised in the used of mentalistic notion [24][25]. The next section will discuss some of the works on ACLs.

3 Related Works

This section discusses some of the works on ACLs in chronological order. Singh [27] discussed the shift of semantic model from mentalistic notion to social interaction. Singh, focused on the possibility the semantic model of mentalistic approach. Agent aren't able uncover the internal state of other agents in computing environment. Hence, the semantic of this model cannot be verified. Thus, it was suggested that the semantic based on social interaction of agent community must be grounded on commitment that is expressed in obligation and prohibition according to society norm. Labrou et al [28] discussed the current landscape of the ACLs the role, origin and concepts of ACLs. KQML and FIPA-ACL were discussed and compared. The application of these ACLs in some of the domain were also elaborated. Kone et al. discussed on the state-of-the-art in ACL [29]. They emphasized on the theory of ACL and discussed some of the pragmatic issue on the implementation of ACLs in the existing models which include KQML, ARCOL, FIPA-ACL, agent oriented programming, open agent architecture, mobile agent communication, and other communication models. Labrou and Finin in another review described pragmatic issues of ACLs such as programming languages, API support, syntax and encoding consideration, services and infrastructures for the ACL and the integration of ACLs with WWW. Steven et al reviewed the issues and challenges of agent communication for open environments [31]. Their review is focused on the agent cities network open environment which is a project in Europe that was used as test bed for agents. Maudet and Chaib-draa provided state-of-the-art in conversational policies and the limitations of this particular approach. These limitations which cover the flexibility and specification were discussed in detail [32]. Chaib-draa and Dignum reviewed the trends in ACLs [33]. They introduced the concept, origin and component of ACLs and discussed the semantic of ACLs in terms of mentalistic approaches and conversational policies. Other important issues such verifications, ontologies and further exploration of semantic of ACLs are also discussed in the paper. Vaniya et al. provided survey on the agent communication language [34] which was mainly focused on semantic and syntax and the implementation of KQML in application.

4 Semantic Model

There are many different semantic models that have been developed for ACLs in order to achieve semantic interoperability. Semantic interoperability allows agent to communicate and understand their communication's message content. There are three primary semantic models that are identified in the development of semantic model for ACL namely mentalistic, conversation policy and social approach. The mentalistic approach defines the semantic of ACL in terms of mental states of agent such as beliefs, desires and intentions. Basically the two dominant ACLs, KQML and FIPA-ACL were developed based on the mentalistic approach in which KQML semantic is based on [15] and FIPA-ACL semantic model is based on [20]. However, the mentalistic approach suffers from the drawback of semantic verifiability which was discussed in [25][27]. Semantic verifiable states that conformance of semantic model could be determined by an independent observer. Since the internal state of agent cannot be uncovered, the conformance of agent towards semantic model cannot be checked. As a result, the semantic interoperability cannot be achieved.

Conversation policy expresses the meaning of the ACL through the composition of speech acts in terms of interaction protocol [35]. Thus, a fixed structure is determined during the adoption of the policy. The implementation of the conversation policy is using finite state. Nevertheless, this approach has two weaknesses which are the lack of flexibility due to the predetermined structure and the lack of well-defined compositional rule the scalability of protocol extension and merging [32].

Social approach defines the ACL semantic in terms of commitments as normative agent society. The effects of the communication acts depend on how the agent should behave in the interaction based on the norm. The concept of commitment is based on the obligation and prohibition of the agent society and is used as the semantic model [36][37]. Obligation is normally specified using deontic logic, however, there are other representations that can be used as well.

4.1 Mentalistic Approaches

[12] discussed the semantic issues of KQML which emphasized on the lack of formal definition of its semantic model. Without a formal semantic, the communication acts are ambiguous and full of confusion. Hence, the ACL is not semantically verifiable and the expected result cannot be predicted. Due to this deficient, Yannis devised a semantic model of pre-condition, post condition and complete condition which are based on mental attitudes [15][16][17]. The semantic model is based on mentalistic notion of beliefs, desire and intention.

Bretier and Sadek presented a rational agent based on the formal theory of interaction called ARTIMIS (Rational Agent Based on Theory of Interaction implemented by a Syntactical Inference Engine) [38]. The communicating agent is modeled as kernel of cooperative spoken dialogue system which model the semantic of communication in first/multi order modal logic of mental attitudes. The reasoning of the communication is based on the inference engine using a theorem prover.

Carron et al. proposed temporal dimension for agent communication language based on speech act theory in mentalistic notion [39]. They modeled the mental states in terms of BDI model with temporal elements which act as constraints for the agent action. The communicative action was modeled in terms of triple consisting of <Pre condition, Post condition, Perlocutionary effect>.

FIPA-ACL was developed by adopting ARCOL agent communication language's semantic model which was based on the semantic of intention of Sadek [21]. The semantic of FIPA-ACL is enhanced by Cohen and Levesque in [20] which is based on quantified multimodal logic which are belief, modal operators for beliefs (B), desires (D), uncertain beliefs (U), and intentions (persistent goals, PG). The semantic of FIPA-ACL is specified in terms of feasibility precondition and rational effect.

Sanjeev presented a group communication semantic for agent communication languages for group interaction [40]. The work derived the semantic of agent communication model based on intention and attempt-based semantics. They treated single-agent communication as a special case of group agent communication.

Although the mentalistic approach has been critiqued due to unverifiable semantic, however, it does lay a solid foundation for agent communication semantic model based on modal logic and possible world semantic model.

4.2 Conversation Policy

Pitt and Mamdani proposed a general semantic framework for ACL in terms of protocol [41]. The protocols are specified in finite state in order to define the context of communication thus, limiting the possibility of the communication act which are applicable in the particular conversational state.

Philips and Link proposed a mechanism that can dynamically combined different conversation policies into conversation specification [42]. The conversation specification allows the contextual issues to be handled during agent communication. Different conversation policies applied to a given conversation specification can change the nature of the interaction.

Nodine and Unruh describe the implementation of conversation policies in InfoSleuth based on finite-state automata [43]. Two mechanisms which are the extension and concatenation were used to simplify the construction of conversation policy. Besides that, a sharable mechanism was introduced to allow the sharing of conversation policy.

Ahn et al. utilized a handshaking mechanism to construct the conversation policy agreement [44]. This approach allowed ad hoc re-implementation of conversation policy in a dynamic changing computing environment.

Despite the disadvantages aforementioned such as rigid structure and format, conversation policy and interaction protocol does give a verifiable semantic model based allowable sequence of message exchanges.

4.3 Social Approaches

Singh presented a social semantics of ACL based on social commitments and temporal logic [26]. The social commitments are based on social context and meta-commitments and are used to capture the legal and social relation between agents. The commitments in semantic model are expressed in terms of deontic concept. Computational tree logic is used to represent the branching time logic in this semantic model.

Colombetti proposed an approach which used agent speech acts and conversations of agent communication language in commitment based approach [45]. This commitment approach is based on social notion. The important components of the commitment based approach are conversational pre commitment and conversational contracts.

Torroni et al proposed an interaction protocol that can be determined by society agent interaction [46]. The semantic model adopted in the communicative act is in terms of commitment which can be expressed in constraints in deontic logic.

Fornara and Colombetti introduced a conditional commitment and precommitment in the social notion based on operational specification within object oriented paradigm [37]. The implementation of operation specification is in an object-oriented approach by the introduction of a commitment class. The conditional commitment is specified with conditional temporal value which at deadline can be active or not. Whereas the pre-commitment will become active after it is accepted by the other agents.

Macro et al presented a logic based social approach communication between societies of agents [47]. There are three important components in the agent society modelling which are the social infrastructure which is responsible for updating the knowledge base, the social organization knowledge base which defines the structure and properties of society such as rules, norms, protocols and social environment knowledge base records the environment data such as events and history. The mental state of the agents in the agent society is defined based on the social effect in terms of obligation and prohibition. Deontic constraints are used to link the events with the obligation and prohibition. Constraint handling rules are used in modelling these constraints.

Federico et al proposed a FIPA compliant goal delegation protocol between agents [48]. It also emphasized that trust element between agents is an important component in goal delegation. Several security methods were proposed for enforcing the trust between agents. New performatives were proposed in this paper for execution of goal delegation protocol. A validation analysis and sample scenario is carried out in order to show the verification of the protocol semantic.

Benoit et al proposed a novel semantics approach for FIPA-ACL based on semantic social attitudes [49]. The social attitudes is represented with communication attitudes based on the concept of grounding [50][51]. In this work, mental attitudes were represented as public commitment instead of private mental state which is not verifiable. Thus, this approach provides a verifiable, formalized and easily adapted model of ACL.

Guido et al introduced a social networks semantic model for ACL semantic model [52]. The intention of the agent in this model is dependent on other agents in the network which is based on dependence network rather than mental attitudes. The advantage of this model is that it is able to model the conversation based on simple graph-theory which in turn can be utilized by the semantic web community.

Social semantic is current trend of semantic model where it provides a verifiable semantic model based on social interaction in terms of notions such as commitment, obligation, prohibition, norms and others. The future research direction will be mostly based on this model where agents in consider social entity that obliged and commit to the computing environment it participates.

4.4 Hybrid Approaches

Boella et al proposed a role-based semantics for agent communication [53]. The novelty of this approach is by embedding both the mentalistic notion and social commitments into the semantic model. In the mentalistic notion, instead of agent's belief and goals, the role of agent's belief and goals is used. The role of the agent is also used for commitment towards the agent society.

Dignum and Linder defined a formal framework for agent communication for social agents [54]. The verifiable model is composed of four components which are the information components for knowledge and belief, action component, motivational component for goal and intentions and social component for commitments and obligation. The model has embodied the mentalistic notion, social commitments and other modal operator together into a formal system.

Nickles et al proposed a semantic model for agent communication in terms of ostensible beliefs and intentions [50]. The difference between ostensible beliefs and mentalistic notion of beliefs and intensions are that the latter is based on introvert agent state which cannot be verified whereas the former is based on the opinion of social structure which can be verified. These opinions can be from individual or groups of agents. A weighted probabilistic approach is used to determine the provenance reliability of the opinion.

5 Conclusion

Agent communication is the key component for agent social interaction. It acts as the medium for the agents to exchange high level knowledge and understand these message in order to achieve the common goal or tasks. This paper provides a review of the ACLs that are commonly used which are KQML and FIPA-ACL. Besides that, in order for the agent to communicate effectively, unambiguous semantic interoperability needs to be achieved. As a result, different semantic models have been developed for ACLs. Yet, among these models there is no one agreeable universal model that is agreed by all. Thus, many endeavours need to be carried to reach the consensus. Nevertheless, several properties can be identified from works that have been done. These properties include verifiable, tractable, decidable, temporal elements and others which are valuable insight for the continuous development of agent communication language.

Acknowledgments. This project is funded by the Ministry of Higher Education, Malaysia under RACE0014-TK-2014.

6 References

1. Genesereth, M. R., Ketchpel, S. P.: Software agents. Communications of the ACM, 37 (7), (1994) 48–53
2. Bradshaw, J.M. (ed.): Software Agents. Cambridge, MA: MIT Press (1997)
3. Nwana, H. S.: Software Agents: An Overview. Knowledge Engineering Review, 11(3), (1996) 205-244.
4. Wooldridge, M.: An Introduction to Multiagent Systems. 2nd Edition. John Wiley & Sons, Inc., New York (2009)
5. Sycara, K.: Multiagent systems. AI Magazine, 19(2), (1998) 79–92

6. Finin, T., J. Weber, G. Wiederhold, M. Genesereth, R. Fritzson, D. McKay, J. McGuire, R. Pelavin, S. Shapiro, C. Beck.: Draft specification of the KQML agent-communication language. Technical report, The ARPA Knowledge Sharing Initiative External Interfaces Working Group, (1993)
7. Finin T., Fritzson R., McKay D., et al.: An Overview of KQML: A Knowledge Query and Manipulation Language. Technical report, Department of Computer Science, University of Maryland, Baltimore County, USA (1992)
8. Neches R., Fikes R.E., Finin T., Gruber T.R., Patil R., Senator T., and Swartout W.R.: Enabling Technology for Knowledge Sharing. AI Mag., 12(3), (1991) 16–36
9. Patil R.S., Fikes R.E., Patel-Schneider P.F., McKay D., Finin T., Gruber T., Neches R.: The DARPA Knowledge Sharing Effort: Progress Report. In: Proc. of Knowledge Representation and Reasoning (1992) 777–788
10. Searle, J.R.: Speech Acts. Cambridge University Press, Cambridge (1969)
11. Austin, J. L.: How to do things with words. Oxford University press (1975)
12. Cohen, P.R., Levesque, H.J.: Communicative actions for artificial agents. In Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95), Menlo Park, California, AAAI Press, (1995) 65–72.
13. Genesereth, M.R. and Fikes, R.E.: Knowledge interchange format, version 3.0 reference manual. Technical Report Logic-92-1, Computer Science Department, Stanford University. (1992)
14. Farquhar, A., Fikes, R. and Rice, J.: The Ontolingua server: a tool for collaborative ontology construction. International Journal of Human-Computer Studies. 46: (1997) 707–727
15. Labrou Y: Semantics for an agent communication language, PhD Thesis dissertation, University of Maryland, Baltimore (1996)
16. Labrou, Y., Finin, T.: Semantics for an agent communication language. In International Workshop on Agent Theories, Architectures, and Languages, Springer Berlin Heidelberg, (1997) 209-214
17. Labrou, Y., Finin, T.: Semantics and conversations for an agent communication language. In: Proceedings of the International Joint Conference on Artificial Intelligence (1997)
18. Chiariglione, L.: FIPA 97 specification, Foundation for Intelligent Physical Agents. (1997)
19. FIPA TC C: FIPA ACL Message Structure Specification. Technical report, IEEE Foundation for Intelligent Physical Agents (2002)
20. Cohen, P. R., Levesque, H.: Persistence, intention and commitment. In Georgeff, M. P. and Lansky, A. L. (Eds.), Reasoning about Actions and Plans: Proceeding of the 1986 Workshop, Morgan Kaufmann, Los Altos, CA, (1986) 297–340.
21. Sadek, M.: A study in the logic of intention. In B. Nebel, C. Rich, & W. Swartout (Eds.), Proceedings third international conference on principles of knowledge representation and reasoning (KR'92), Morgan Kaufmann Publishers. (1992) 462–473
22. FIPA TC C: Fipa communicative act library specification. Technical report, IEEE Foundation for Intelligent Physical Agents (2002)
23. FIPA TC C: Fipa SL content language specification. Technical report, IEEE Foundation for Intelligent Physical Agents (2002)
24. http://www.fipa.org/
25. Wooldridge, M.: Verifiable semantics for agent communication languages. In: International Conference on Multi-Agent Systems (ICMAS 1998), Paris, France (1998)
26. Singh, M.P.: A social semantics for agent communications languages, in Proceedings of the IJCAI-99 Workshop on Agent Communication Languages, F. Dignum, B. Chaib-draa, and H. Weigand, eds., Berlin: Springer-Verlag (2000)

27. Singh, M.P.: Agent communication languages: Rethinking the principles. IEEE Computer 31(12), (1998) 40–47
28. Labrou, Y., Finin, T., Peng, Y.: Agent communication languages: The current landscape. IEEE Intelligent systems, 14(2), (1999) 45-52
29. Kone, M.T., Shimazu, A., Nakajima, T.: The state of the art in agent communication languages. Knowledge and Information Systems 2, (2000) 259–284
30. Labrou, Y., Finin, T.: History, State of the Art and Challenges for Agent Communication Languages. INFORMATIK - Zeitschrift der schweizerischen Informatik organisationen 7, (1999) 17–24
31. Willmott, S., Dale, J., and Charlton, P.: Agent Communication Semantics for Open Environments: Issues and Challenges (No. EPFL-REPORT-52461). (2002)
32. Maudet, N., Chaib-draa, B.: Commitment-based and Dialogue-game based Protocols– News Trends in Agent Communication Language. The Knowledge Engineering Review 17, (2002) 157–179
33. Maudet N., Chaib-draa B.: Trends in agent communication language. Comput. Intell. 18(2), (2002) 89–101
34. Vaniya, S., Lad, B., Bhavsar, S.: A Survey on Agent Communication Languages. In 2nd International Conference on Innovation, Management and Service (ICIMS)-Singapore (2011)
35. Greaves, M., Holmback, M., Bradshaw, J.: What is a conversation policy? In: Dignum, F.P.M., Greaves, M. (eds.) Issues in Agent Communication. LNCS, vol. 1916, pp. 118–131. Springer, Heidelberg (2000)
36. Yolum, P., Singh, M.: Commitment machines. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, pp. 235–247. Springer, Heidelberg (2002)
37. Fornara, N., Colombetti, M.: Operational specification of a commitment-based communication language. In: Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), Bologna, Italy (2002)
38. Breiter, P., Sadek, M.D.: A rational agent as a kernel of a cooperative dialogue system: Implementing a logical theory of interaction. In: ECAI 1996 Workshop on Agent Theories, Architectures, and Languages, pp. 261–276. Springer, Heidelberg (1996)
39. Carron, T., Proton, H., Boissier, O.: A Temporal Agent Communication Language for Dynamic Multi-Agents Systems, Proceedings of 9th MAAMAW, LNAI 1647. (1999)
40. Kumar, S., Huber, M.J., McGee, D., Cohen, P.R., Levesque, H.J.: Semantics of agent communication languages for group interaction. In: Proceedings of the 17th Int. Conf. on Artificial Intelligence, Austin, Texas, (2000) 42–47
41. Pitt, J., & Mamdani, A.: A protocol-based semantics for an agent communication language. In Proceedings of the international joint conf. on artificial intelligence IJCAI (1999) 486–491
42. Phillips, L.R., Link, H.E.: The role of conversation policy in carrying out agent In F. Dignum & M. Greaves, (Eds.), Issues in agent communication, volume 1916 of Lecture Notes in Computer Science. Springer (2000)
43. Nodine, M., Unruh, A.: Constructing Robust Conversation Policies in Dynamic Agent Communities. In: Dignum, F., Creaves, M. (eds.) Issues in Agent Comm., Springer, Heidelberg (2000)
44. Ahn, M., Lee, H., Yim, H., Park, S.: Handshaking Mechanism for Conversation Policy Agreements in Dynamic Agent Environment. In: First International Joint Conference on Autonomous Agents and Multi-Agent Systems (2002)

45. Colombetti, M.: A commitment-based approach to agent speech acts and conversation. In: Proc. Workshop on Agent Languages and Communication Policies, 4th International Conference on Autonomous Agents (Agents 2000), Barcelona, Spain, (2000) 21–29
46. Torroni, P., Mello, P., Maudet, N., Alberti, M., Ciampolini, A., Lamma, E., Sadri, F., Toni, F.: A logic-based approach to modeling interaction among computees (preliminary report). In: UK Multi-Agent Systems (UKMAS) Annual Conference, Liverpool, UK (2002)

Simulation and Fabrication of Micro Magnetometer Using Flip-Chip Bonding Technique

Tengku Muhammad Afif bin Tengku Azmi, Nadzril bin Sulaiman

International Islamic University Malaysia, Faculty of Engineering, Gombak P.O. Box 10,
50728 Kuala Lumpur, Malaysia
afif5785@gmail.com, nadzril@iium.edu.my

Abstract. Magnetic field detection has been widely accepted in many applications such as military systems, outer space exploration and even in medical diagnosis and treatment. Low magnetic field detection is particularly important in tracking of magnetic markers in digestive tracks or blood vessels. The presence of magnetic fields' strength and direction can be detected by a device known as magnetometer. A magnetometer that is durable, room temperature operation and having non-movable components is chooses for this project. Traditional magnetometer tends to be bulky that hinders its inclusion into micro-scaled environment. This concern has brought the magnetometer into the trend of device miniaturization. Miniaturized magnetometer is usually fabricated using conventional microfabrication method particularly surface micromachining in which micro structures are built level by level starting from the surface of substrates upwards until completion of final structure. Although the miniaturization of magnetometer has been widely researched and studied, the process however is not. Thus, the process governing the fabrication technique is studied in this paper. Conventional method of fabrication is known as surface micromachining. Besides time consuming, this method requires many consecutive steps in fabrication process and careful alignment of patterns on every layer which increase the complexity. Hence, studies are done to improve time consuming and reliability of the microfabrication process. The objective of this research includes designing micro scale magnetometer and complete device fabrication processes. A micro-scale search coil magnetometer of 15 windings with 600µm thickness of wire and 300µm distance between each wire has been designed.

Keywords: Magnetometer, microfabrication, miniaturization, micro-scale.

1 Introduction

Magnetometer is a device used to detect external magnetic field. Basic magnetometer contains only a thin layer of metal that will detect the change in voltage when placed in magnetic field [1]. Common magnetometers are bulky. MEMS has brought an advancement to the development of magnetometer which are small, light, low power consumption, high sensitivity and high resolution [2]. Basic types of magnetometer are scalar and vector magnetometer. Scalar measured the magnitude of the vector magnetic field while the vec-

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_48

tor magnetometer measures the vector components of a magnetic field [3]. A vector is a mathematical entity with both magnitude and direction. The Earth's magnetic field at a given point is a vector. A magnetic compass is designed to give a horizontal bearing direction, whereas a vector magnetometer measures both the magnitude and direction of the total magnetic field. Three orthogonal sensors are required to measure the components of the magnetic field in all three dimensions. The advancement in technology lead to the miniaturization of magnetometer. Some common and actively research micro magnetometer are SQUID, ferromagnetic and magnetorisistor. The early develop micro SQUID was introduced by Mark Ketchen at IBM [4]. The need to miniaturized high sensitivity low power magnetometer has led to the development of ferromagnetic MEMS magnetometer [5]. This new micro magnetometer has minimum the power intake and maintain its sensitivity although being scaled-down. This device consists of dual 4.2×2×200 μm3 polysilicon torsion bars and a 100×100×13.4 μm3 ferromagnetic plate.

2 Background

Search coil can be miniaturized by means of MEMS technology. Its components include, coil and core (air). When exposed to an external magnetic field, the micro coil will generate a voltage different and it represent the magnitude of magnetic flux density. When measuring the magnetic flux density, which is in Tesla (T), a voltmeter (v) is used as a representation.

A planar design of search coil magnetometer is used in this work. The main reason is to simplify the fabrication process. Figure 1 shows micro coil magnetometer while figure 2 shows the side view. The design for both driving and sensing coil are made identical to each other with different orientation. It will then apply the flip chip bonding technique.

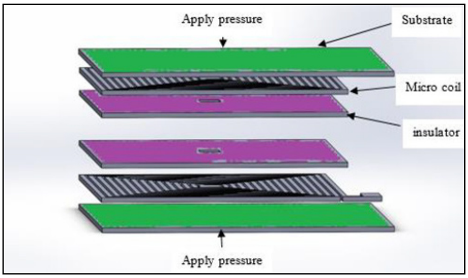


Fig. 1. Micro coil magnetometer

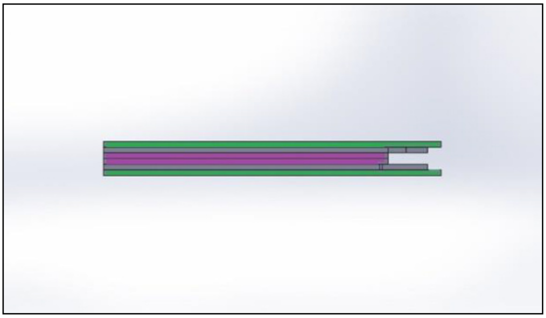


Fig. 2. Side view

Search coil magnetometer can detect a magnetic field as low 20fT and no upper limit [6]. It works based on the principal of faraday's law of induction. It is suitable to use in detecting low and high magnetic field. The advantage over fluxgate magnetometer meter is that it consumes less power. Moreover, the design of search coil is simpler compare to fluxgate. Hence, it will make it easier to focus on the fabrication technique part which is the main idea in this paper. Search coil is a type of magnetometer that operates based on law of voltage induction. This law was introduced by Faraday which states a change in magnetic fields could produce an electromotive force or voltage. This voltage will in turn produce electric currents in a closed circuit. It is capable to produce magnetic fields in the range of 20 femtoTesla up to unlimited upper range. It is a simple and low cost magnetic field sensor. The downside is that it is unable to measure static or slow varying magnetic fields. Figure 3 shows the comparison of magnetometer in term of sensitivity and frequency. It shows that the coil-based (search coil) can detect the low magnetic field at high frequency.

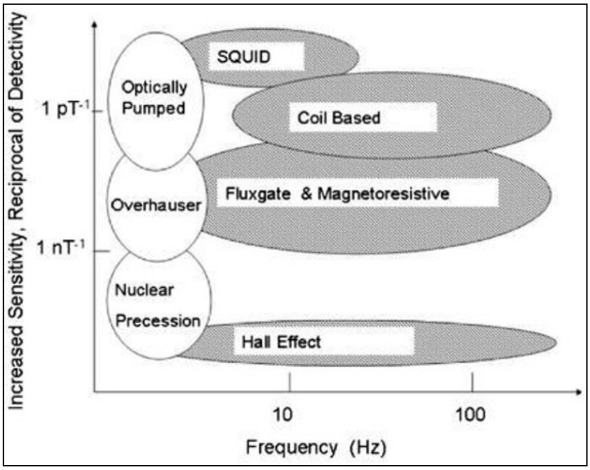


Fig. 3. Comparison of sensitivity and frequency of different type of magnetometer [7]

3 Design and Simulation

Planar design of the search coil magnetometer with the help of SolidWorks was chosen. Planar design eases the fabrication of micro coil. The design consideration and specifications were done based on the simulation obtained through COMSOL 5.0 software. Here, the strength of magnetic field was compared for different thickness and different spacing.

The most important aspect to be considered while designing a magnetometer is the coil. Thus, to ensure the design micro coil meet the requirements, several simulations must be done. Figure 4 shows simulation for different thickness (20µm and 70µm) of coil with the same spacing (70µm) between each coil. Figure 5 shows simulation for different spacing (20µm and 70µm) between coil with the same thickness (40µm).

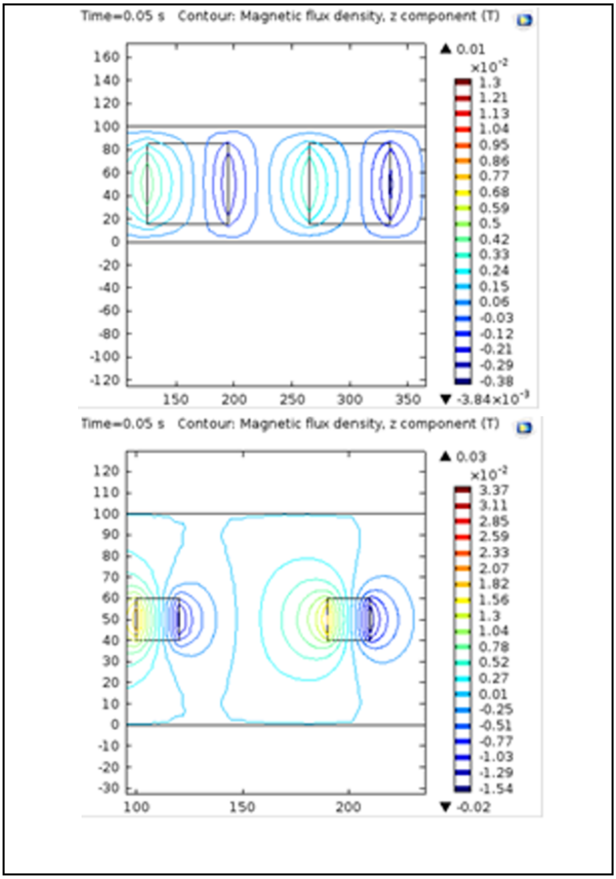


Fig. 4. Different thickness same spacing

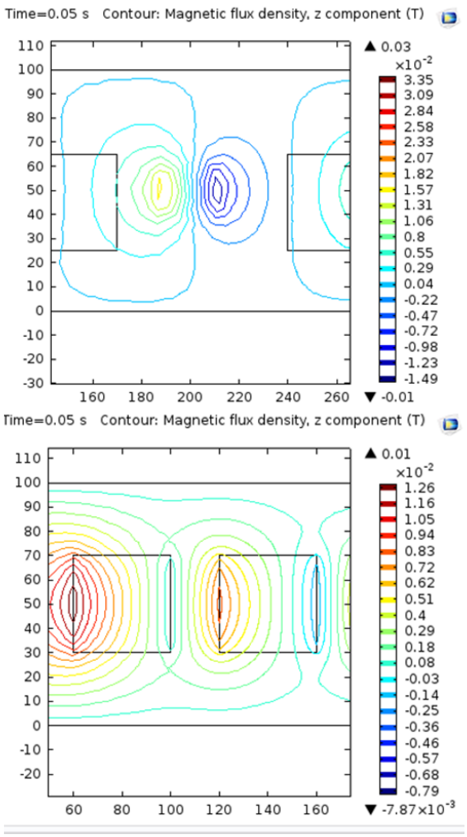


Fig. 5. Same thickness different spacing

These results are then compared with the mathematical equation of magnetic field shown in equation (1). Based on the results obtained, a thickness of 60μm with spacing of 40μm is design as the optimum design of micro coil with 30 turns. We kept the number of turns minimum to prevent the magnetometer from being oversized (from micro). Air represents the core for search coil. Saving the time and sources to fabricate it.

$$B=\mu 0I/Area \qquad (1)$$

4 Result

4.1 Fabrication of Search Coil

The idea of this part is to determine the fastest possible technique that can be achieved to fabricate a micro magnetometer. A conventional way of fabrication depends on surface micro machining technique only. Here, surface micro machining is combined with the flip-chip based technique to shorten the time taken. Besides, it could increase the possibility of refabricating and minimizing the used of substances.

Surface Micromachining. The sequence in Fig. 6 shows the steps taken to complete the fabrication of the magnetometer using surface micromachining with flip chip bonding technique. It takes nine steps to complete the fabrication process provided both coils are fabricated in line with each other. Not only that the steps are lesser than the process which depends on surface micromachining only, it is easier to redo the process if there are mistakes along way.

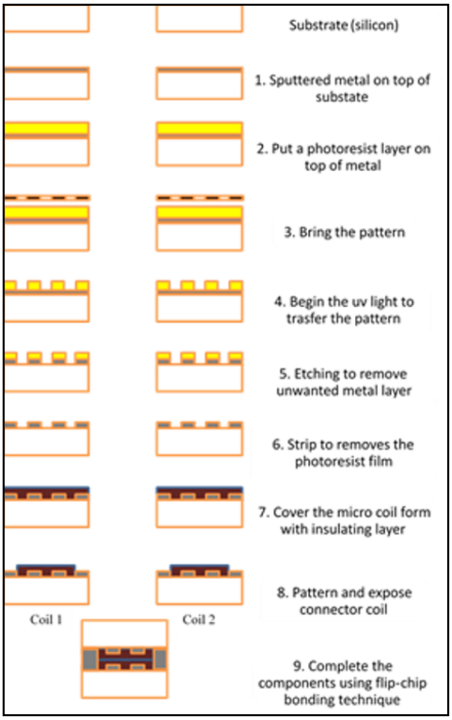

Substrate (silicon)
1. Sputtered metal on top of substate
2. Put a photoresist layer on top of metal
3. Bring the pattern
4. Begin the uv light to trasfer the pattern
5. Etching to remove unwanted metal layer
6. Strip to removes the photoresist film
7. Cover the micro coil form with insulating layer
8. Pattern and expose connector coil
Coil 1
Coil 2
9. Complete the components using flip-chip bonding technique

Fig. 6. Surface Micromachining + Flip-Chip Bonding

Surface Micromachining + Flip Chip Bonding. Fig. 7 shows the steps taken to complete the fabrication of magnetometer using surface micromachining only. As shown, it takes fourteen steps which is five steps more than the previously shown. The process cannot proceed unless the previous process is done. This will prolong the duration and steps compared to previous technique which enable the process to be done simultaneously. Besides, it is much more complicated to correct mistakes done during the process since the steps are depending on each other.

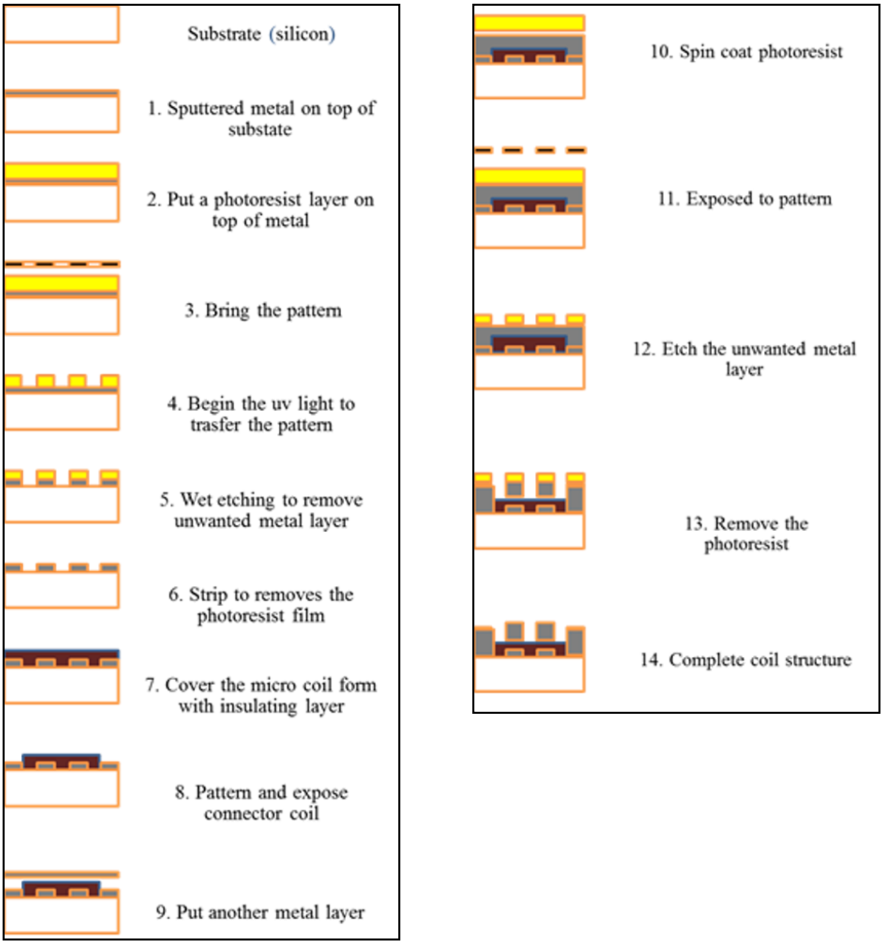

Substrate (silicon)
1. Sputtered metal on top of substate
2. Put a photoresist layer on top of metal
3. Bring the pattern
4. Begin the uv light to trasfer the pattern
5. Wet etching to remove unwanted metal layer
6. Strip to removes the photoresist film
7. Cover the micro coil form with insulating layer
8. Pattern and expose connector coil
9. Put another metal layer
10. Spin coat photoresist
11. Exposed to pattern
12. Etch the unwanted metal layer
13. Remove the photoresist
14. Complete coil structure

Fig. 7. Surface Micromachining

The steps taken to fabricate micro search coil magnetometer begin from the preparation of substrate until the flip-chip bonding are as follow. Note that Fig. 8 and Fig.9 show the image taken during the process

1) Prepare the substrate (silicon).
2) Insert the substrate into Physical Vapor Deposition (PVD) machine to get a surface of metal (aluminum) on the substrate.
3) Clean the aluminum surface with DI water to remove unwanted object.
4) Start putting positive photoresist on top of aluminum and spin coat to get evenly distributed surface.
5) Soft bake to harden a bit the photoresist.
6) Expose pattern from transparency mask.
7) Develop the pattern in developer solution and rinse with Deionized (DI) water once the pattern is visible with naked eye. Spin coat to dry.

8) Hard baked to make sure the patterned photoresist is well preserved before putting in aluminum etch solvent.
9) Put in aluminum etch solvent and rinsed with DI water once the unwanted aluminum has been etch out. Spin coat to dry.
10) Used acetone to remove the remaining photoresist on aluminum pattern. Spin coat.
11) Insulate and combined the two coil (flip-chip).

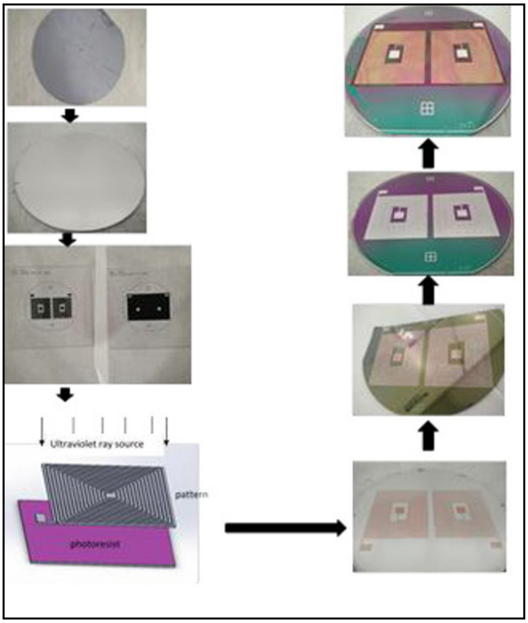


Fig. 8. Fabrication of Search Coil

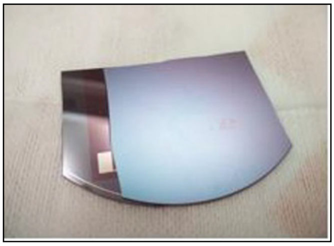


Fig. 9. Complete Prototype

4.2 Search Coil Magnetometer Functionality Testing

After completing the fabrication of micro-scaled search coil magnetometer, the prototype is tested to test its functionality. For calibration purposed, the search coil magnetometer output is compared with the output from gauss meter. The experimental setup equipments are helmholtz coil (EM 6723), gauss meter (410), power supply (GPS-3303), search coil magnetometer, Digital Multimeter (DMM) and the holder. The results obtained shown in Table 1.

Table 1. Output from gaussmeter and fabricated magnetometer

Distance (cm)	Voltage (V)	Current (A)	Magnetic field (mT)	Induced emf (mV)
20	3	0.08	-0.07	No significant changes
	6	0.16	-0.05	
	9	0.24	-0.02	
	12	0.32	0	
	15	0.41	0.07	
	18	0.49	0.10	
	21	0.57	0.15	
	24	0.65	0.18	
16	3	0.08	-0.13	-0.1
	6	0.16	-0.16	-0.2
	9	0.24	-0.18	-0.2
	12	0.32	-0.20	-0.2
	15	0.41	-0.23	-0.3
	18	0.49	-0.25	-0.3
	21	0.57	-0.28	-0.4
	24	0.65	-0.30	-0.5
10	3	0.08	-0.07	-0.3
	6	0.16	-0.05	-0.6
	9	0.24	-0.02	-0.8
	12	0.32	0	-1.3
	15	0.41	0.05	-1.7
	18	0.49	0.07	-1.9
	21	0.57	0.10	-1.9
	24	0.65	0.14	-1.9

5 Conclusion

Micro fabrication is a wide and interesting area to study. Many of new things can still be found in fabrication. As for micro magnetometer fabrication, although the things that can be explored are limited, yet there's a small fraction of untouched topic in the micro fabrication technique. This paper has touched the small untouched topic that yet to be discovered. Throughout the research, the time needed to complete the micro fabrication technique compared to the conventional surface micromachining used has been improved. Besides, the re-work potential of the prototype is also possible. Thus, the overall objectives have been achieved. The objective about implementing a new technique as opposed to the conventional surface micromachining has been achieved. The combination of flip-chip bonding technique together with the surface micromachining improves the fabrication time and reduces the process complexity. Conventional fabrication which depends on layer by layer technique can now be done in a different way. The micro-scale search coil fabricated can function as it is supposed to. Detecting magnetic flux density and produce an induced voltage as an output.

Acknowledgments. The authors acknowledge the supports from the Ministry of Higher Education and the International Islamic University Malaysia under grant no. RAGS14-033-0096.

References

1. B. Y. Y. Cai, Y. Zhao, X. Ding, and J. Fennelly, “Magnetometer basics for mobile phone applications,” no. February, 2012.
2. A. L. Herrera-may, L. A. Aguilera-cortés, P. J. García-ramírez, N. B. Mota-carrillo, W. Y. Padrón-hernández, and E. Figueras “Development of Resonant Magnetic Field Microsensors : Challenges and Future Applications,” 2011.
3. Edelstein, Alan (2007). "Advances in magnetometry". J. Phys.: Condens. Matter 19: 165217 (28pp).
4. Wikswo, J. P. 1995. Squid magnetometers for biomagnetism and non-destructive testing: important questions and initial answers. IEEE Transactions on Applied Superconductivity 5(2): 74-120.
5. Yang, H. H., Myung, N.V., Yee, J., Park, D-Y., Yoo, B-Y., Schwartz, M., Nobe, K. & Judy, J.W. ferromagnetic micromechanical magnetometer. Sensors and actuator A 97-98: 88-97, 2002.
6. James Lenz, J., & Edelstein, A. s. magnetic sensors and their applications. IEEE Sensors Journal 6(3): 631-649, 2006.
7. A. S. Edelstein, J Burnette, G. A. Fischer, S. F. Cheng, W. F. Egelhoff, Jr., P. W. T. Pong, R. D. Mcmichael, E. R. nowak “advance in magnetometry thriugh minituarization”, 2008

A Review on Recognition-Based Graphical Password Techniques

Amanul Islam[1], Lip Yee Por[1], Fazidah Othman[1], Chin Soon Ku[2]

[1]University of Malaya, Kuala Lumpur 50603, Malaysia
[2]University of Tunku Abdul Rahman, Kampar 31900, Malaysia,
aman.um16@siswa.um.edu.my, porlip@um.edu.my, fazidah@um.edu.my,
kucs@utar.edu.my

Abstract: This paper reviews the recognition-based graphical password system. Twenty-five recognition-based graphical password systems are studied and analyzed with regards to their security threats. Countermeasures and suggestions are given to prevent and reduce the security threats. A comparison summary of the selected recognition-based graphical password system is deliberated at the end of this paper.

Keywords: Graphical, Password, Authentication, Recognition, Method, System

1 Introduction

With accelerated the evolution of systems and applications, the urge for a potent computer security is growing [1].The majority of the computer systems and applications are preserved with user identification & authentication. However, many of them are having flaws due to an acquiescent and proficient user. Although, there are many ways to authenticate a person, the most commonly used means of authentication method is using the password method.

Passwords always comply two fundamental contradicted requirements where they must be secure and easy to remember [2]. However, this is hard to achieve using alphanumeric passwords because a long and random password is secure but it is hard for the users to remember. Therefore, most of the users tend to use weak password [3]. Graphical password was then introduced as an alternative authentication method for alphanumeric passwords to overcome the memorability issue [3].

Back in 1996, Greg Blonder first explained the concept of graphical passwords [4]. Graphical password is easier to remember than alphanumeric password, which is an important advantage of it [4]. Graphical passwords utilize images in place of the alphanumeric passwords since humans are readily able to recognize images than a series of characters [5]. Human beings have the capability to recognize places they visit, other people's faces, and things [6]. Therefore, graphical password system paves a path by presenting a lot easier to use passwords whilst enhancing the security level [7]. Except for these improvements, the most acclaimed issue with graphical passwords is the shoulder-surfing attack [3]. Shoulder-surfing leads to employing direct observation methods, for instance, eyeballing over someone's shoulder, to obtain

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_49

information [3]. Numerous researchers have endeavored to resolve this obstacle by giving distinct procedures. Hence, we started this research to find out the problem of algorithmic level of different recognition-based graphical password schemes how they implemented these schemes and why the problem of shoulder-surfing and other attacks arise in the field of recognition-based graphical passwords.

2 Methodology

This research begins with gathering information about existing recognition-based graphical password systems. The information is amassed from different sources, for example – journals, conference papers, and legitimate sites. The perceived systems are dissected to discover their qualities and inadequacies. Results from the investigation permit a prevalent perception of the present issues and troubles impacting existing graphical password systems. This information is used as a piece of the way toward making and specifying the research objectives.

3 Research Background

A graphical password system is a system that uses objects (images/icons/symbols) to perform authentication [8]. There are two main procedures in a graphical password system – enrolment procedure and authentication procedure. In the enrolment procedure, users are required to register certain objects from a database as their password [9]. In the authentication procedure, the users are given a challenge set to perform authentication. The users are required to identify the correct objects before they can access into a secure system.

Graphical password can be categorized into three categories – recognition-based, recall-based, and cued recall-based [10]. Recognition-based graphical password systems generally require users to register and memorize objects during enrolment procedure. The users are compelled to click the correct objects during the authentication procedure. The correct objects in each challenge set can be the registered objects, part of the registered objects or pass-objects that identify using certain methods. In recall-based graphical password systems, based on the registered objects, users are needed to remember and portray a covert drawing within a given grid or a blank canvas. On the other hand, cued recall-based graphical password systems needed users to remember and pinpoint target on specific locations within a picture.

In this paper, we only focus on the study of recognition-based graphical password systems because based on our reviewed, majority of the articles were found belongs to this category. The selected recognition-based graphical password systems are reviewed as below.

4 Recognition-Based Graphical Password Systems

Passfaces^{TM} is a commercial product and it is one of the earliest recognition-based graphical password systems introduced by Passfaces^{TM} Corporation [11]. During the enrolment procedure, users are required to register pictures of human faces. In the authentication procedure, the users are requested to click on the registered pictures to login. This system is simple and easy to use [12]. However, Passfaces^{TM} is vulnerable to direct observation shoulder-surfing attack. Moreover, users who have prosopagnosia syndrome (face blindness) will find this system difficult to use.

In Déjà Vu system, users are needed to register several "random art" images during the enrolment procedure [1]. In the authentication procedure, the users are required to click on the registered images to login. This system is simple and easy to use. However, the direct selection of the pictures during authentication allows direct observation shoulder-surfing attack to be done successfully.

The Picture Password system, proposed by [13], is designed for mobile devices like the PDA. Users can either choose images from one of three predefined themes or provide their own images during enrolment procedure. In the authentication procedure, users are needed to click on the registered images to login. This system tries to increase the password space by allowing two images to be selected as one image. However, the direct selection of the pictures during authentication allows direct observation shoulder-surfing attack to be carried out.

Story system utilizes the same enrolment and authentication methods as in Passfaces^{TM} system [14]. Despite using human faces images, this system uses non-human images. However, this system also suffers from direct observation shoulder-surfing attack.

In Triangle system, users are needed to register and remember three icons during the enrolment procedure [1]. In the authentication procedure, the users are required to form a polygon using the three registered icons virtually. The users need to click one of the icons (pass-icon) within the polygon area (convex hall) to complete a challenge set. The users are required to pass several challenge sets before they can login. This system uses other icons besides the registered icons to login. Thus, the system is able to resist direct observation shoulder-surfing attack.

In Moving Frame system [8], users are required to register and remember three icons during the enrolment procedure. In the authentication procedure, the users need to rotate the frame to ensure the two registered icons, which located within the frame, can form a straight line. The users are required to pass several challenge sets before they can login. This system does not required users to click on the registered icons. Therefore, it is able to resist direct observation shoulder-surfing attack. However, there are only four ways for the users to rotate the frame. Thus, chances for attackers to guess the correct rotation are quite high.

In Special Geometric Configuration (SGC) system [8], users are required to register four icons during the enrolment procedure. In the authentication procedure, users need to locate the registered icons. Then, the users need to use two of the registered icons to virtually form a line. The users need to click on the intersection icon that made by the two virtual lines to login. Similarly, this system does not required users to click on the registered icons. Therefore, the system is able to resist direct observation shoulder-surfing attack.

In Scalable Shoulder-Surfing Resistant Textual-Graphical Password (S3PAS) system [15], users are required to register at least three images during the enrolment procedure. In the authentication procedure, the users have to mentally construct a triangle using a group three characters, and then click on any character within the area of the virtual triangle formed. The process will be repeated for all possible groupings. For example, if a user's registered "L0V3", the possible groupings are, "L0V", "0V3", "V3L" and "3L0". Similar to the triangle system, this system is able to resistant direct observation shoulder-surfing attack because it does not use the registered images to login.

Visual Identification Protocol (VIP) version one and version two are two systems that predefined a set of registered images to the users instead of allowing the users to register themselves during the enrolment procedure [16]. In the authentication procedure, the users are required to identify the correct images in sequence before they can login. The difference between VIP1 and VIP2 is the arrangement of the pictures and the number of pictures used. VIP1 uses ten pictures and the arrangement of the picture is similar to the arrangement of keypad numbers in an ATM machine. On the other hand, VIP2 uses 3 x 4 grid cell interface to perform user authentication. These systems are simple and easy to use. However, the registered pictures chosen by the users can be shoulder-surfed easily as well. Therefore, both systems are vulnerable to direct observation shoulder-surfing attack. In the VIP version 3 [16], users are needed to register eight pictures during enrolment procedure. In the authentication procedure, only four of the registered pictures will be shown in a 4 x 4 grid cell. The rest of the grid cells are filled with decoy pictures. To login, the uses are required to click on the registered pictures in sequence. This system can reduce direct observation shoulder-surfing attack although the attackers can shoulder-surf the registered pictures clicked by the users every time they login. The main reason this system can reduce direct observation shoulder-surfing attack is because in every challenge set, only part of the registered pictures is shown. Therefore, it will take time and extra effort for the attackers to analyze the correct registered pictures used by the users.

Use Your Illusion system utilizes the same enrolment and authentication methods as in Passfaces^{TM} system [17]. Despite using human faces images, this system uses distorted images. Although the distorted pictures are hard to be seen clearly, attackers can still shoulder-surf the clicked pictures. Thus, this system is vulnerable to direct observation shoulder-surfing attack. Moreover, this system suffers from a small password space.

In ColorLogin system [18], users are needed to choose a color and a set of icons in the enrolment procedure. The users can use the registered color as background to help them find their registered icons. In the authentication procedure, users are required to click on the rows that contain the registered icons in an N x N grid cell. Once clicked; the entire row will be locked. All the affected icons will change to a "lock" icon. To complete a challenge set, the users have to ensure all the registered icons are locked. The users have to perform several challenge sets in order to login. ColorLogin system is able to reduce direct observation shoulder-surfing attacks because the registered icons are not chosen directly during the login process. However, attackers can still shoulder-surf the row that clicked by the users. Moreover, this system is vulnerable to guessing attacks due to small password space.

Graphical Password with Icons (GPI) and Graphical Password with Icons suggested by the System (GPIS) are proposed by [19]. In GPIS, users are required to register six icons during the enrolment procedure. In GPIS, the six icons are assigned to the users during the enrolment procedure. In the authentication procedure, both systems required the users to identify and click on the registered/assigned icons among 150 icons to login. Therefore, both systems are vulnerable to direct observation shoulder-surfing attack because the registered/assigned icons chosen by the users can be easily shoulder-surfed.

There are two variations of What You See is What You Enter (WYSWYE) system [20]. In both variations, users are required to register four images during the enrolment procedure. For the first variation, called the Horizontal Reduce Scheme (HRS), users are presented with a 7x4 grid during the authentication procedure. The users have to find the columns and mentally eliminate columns that do not have their registered images. The result will be an Nx4 grid with the maximum size being a 4x4 grid. The users then need to key in the corresponding position of the registered images in the password input grid. The second variation, called the Dual Reduce Scheme (DRS), users are presented with a 5x5 grid. The users have to eliminate a row and a column that does not have their registered images. The result will be an M x N grid, again with the maximum size being a 4x4 grid. Similar to the first variation, users are required to key in the position of the registered images in the password input grid. WYSWYE-HRS and WYSWYE-DRS are able to reduce direct observation shoulder-surfing attack because the registered images are not selected during the authentication processes. However, attackers can still shoulder-surf the value of keyed in by the users and map with the position of the registered images.

In Por's system [21], users are required to register eight images in the enrolment procedure. In the authentication procedure, the users are required to click four or five registered images to login. Similar to VIP3, this system can reduce direct observation shoulder-surfing attack because only part of the registered images is used for every challenge set.

In Manjunath's system [22], users are required to register a string (8 to 15 characters) and choose one color (eight colors are given) during the enrolment procedure. In the authentication procedure, eight color sectors are shown and each sector is filled with eight random characters. To login, the users are required to move the registered color sector to the registered characters. This system can prevent direct observation shoulder-surfing attack because the registered string and color are not directly used.

In Haque's scheme [23], users are required to register their username and at least several images during the enrolment procedure. After that, a set of questions will be given to the users. The users need to pair each question with three registered images. In the authentication procedure, the users are required to recognize the correct images based on the question asked. This system is easy and simple to use. However, this system cannot prevent direct observation shoulder-surfing attack because the direct selection of the registered images during an authentication process can be easily observed and shoulder-surfed.

In Pooja system [24], users are required to register several images during enrolment procedure. During the authentication procedure, the users are required to identify the registered images from the 4 x 4 grid cell. This system is simple and easy to use. However, this system is vulnerable to direct observation shoulder-surfing attack because the direct selection of the registered images during an authentication process can be easily observed and shoulder-surfed.

In CuedR system [25], users are required to register six animal images during the enrolment procedure. In the authentication procedure, the users are required to key in the character associated with the registered images in sequence. This system is vulnerable to direct observation shoulder-surfing attack because attackers can decompose the password string then associated each character with the unique animal image in a challenge set.

In Digraph Substitution Rules (DSR) system [3], users are needed to register a username and register two images in the enrolment procedure. In the authentication procedure, users need to click on a pass-image based on the registered images and the three digraph substitution rules. The users have to complete several challenge sets before they can login. This system can prevent direct observation shoulder-surfing attack because the users will never click on the registered images.

In WordPassTile system [26], users are required to register five Tiles (a unique word) in the enrolment procedure. In the authentication procedure, users are required to click on the Tiles provided in a specific sequence. This system is vulnerable to direct observation shoulder-surfing attack because the direct selection of the Tiles during an authentication process can be easily observed and shoulder-surfed.

In Graphical-Text Password Authentication (GTPA) system [27], users are required to register four images in the enrolment procedure. In the authentication procedure, the users have to click on the first pass-image within a 10 x 10 grid cell based on the pair of numbers associated with the first registered image. After clicking, the images and the pair of numbers will be re-shuffled using uniform randomization algorithm. The user then has identified the second pass-image based on the pair of numbers associated with the second registered image. The same process keeps repeating until the users click on the fourth pass-image before the user can login. This system can prevent direct observation shoulder-surfing attack because the images clicked by the users could be the registered image or the decoy image.

5 Common attacks in recognition-based graphical password system

The following are the common security threats for recognition-based graphical password systems: -

Guessing attack – It is the process of getting the password of a user by predicting or resolving the password [1, 28]. Most of the recognition-based graphical password systems, which have small password space usually, will encounter such security threat. There are several ways to overcome such attack. For example, increase the password space, use partial registered objects (images/icons/symbols) or pass-objects (pass-images/pass-icons/pass-symbols) to login.

Direct observation attack – It is a type of shoulder-surfing attack for example eyeballing over someone's shoulder to obtain information [3]. Most of the recognition-based graphical password systems, which uses direct registered objects, will encounter such security threat. To overcome or reduce this attack, indirect objects for example pass-objects can be used to login.

Frequency of Occurrence Analysis (FOA) attack – It only happens in recognition-based systems that use uniform randomization algorithm to perform selection [21]. Due to the fact that the sampling size of the registered objects is relatively smaller than the decoy objects sampling size, when uniform randomization algorithm is used, the probability the registered objects will always appear in a challenge set while the same distracter image will only appear occasionally in every challenge set [21]. To overcome or reduce this attack, an authentication system can use fix objects or prevent using large amount of decoy objects in every challenge set.

6 Result and Discussion

Table 1. Recognition-based graphical password and its security threats

Graphical Password Schemes	Direct observation attack	FOA	Guessing attack
Passfaces^{TM}	✗	✗	✗
Déjà Vu	✗	✗	✗
Picture Password system	✗	✗	✗
Story	✗	✗	✗
Triangle system	✓	N/A	✓
Moving Frame system	✓	N/A	✗
SGC	✓	N/A	✓
S3PAS	✓	N/A	✓
VIP1	✗	N/A	✗
VIP2	✗	N/A	✗
VIP3	can reduce	✗	✓
Use your illusion	✗	✗	✗
ColorLogin	✓	✗	✗
GPI	✗	N/A	✓
GIPS	✗	N/A	✓
WYSWYE-HRS	can reduce	N/A	✓
WYSWYE-DRS	can reduce	N/A	✓
Por’s system	can reduce	can reduce	✓
Manjunath’s system	✓	N/A	✓
Haque’s system	✗	N/A	✓
Pooja’s system	✗	✓	✓
CuedR	✗	N/A	✓
DSR	✓	✓	✓
WordPassTile	✗	✓	✓
GTPA	✓	✓	✓

Note: ✗ = vulnerable to the attack

✓ = invulnerable to the attack
N/A=not applicable

Table 1 shows the comparison table among the reviewed system. From the table, majority of the reviewed systems are vulnerable to direct observation shoulder-surfing attack. Only systems that used partial objects to login can reduce such attack. For example, VIP3, WYSWYE-HRS, WYSWYE-DRS, and Por's system. Systems that use indirect input or pass-objects instead of the registered objects to login can prevent this attack. Examples of these systems are – Triangle system, Moving Frame system, SGC, S3PAS, ColorLogin, Manjunath's system, DSR, and GTPA.

In terms of FOA attack, there are only few systems get affected because these systems used uniform randomization to perform selection. For example Passfaces^{TM}, Déjà Vu, Picture Password system, Story, Photographic authentication, VIP3, Use your illusion and ColorLogin. There are few systems are able to resist such attack because they used fix number of objects every time to login. Examples of these systems are – Pooja's system, DSR, WordPassTile and GTPA. Other systems are not relevant because they are not using uniform randomization to perform selection.

In terms of FOA attack, there are only few systems get affected because these systems have small password spaces. Example of these systems are – Passfaces^{TM}, Déjà Vu, Picture Password system, Story, Moving Frame system, VIP1, VIP2, Use your illusion and ColorLogin.

7 Conclusion

In this study, several specific security threats such as guessing attack, direct observation and FOA that encountered by recognition-based graphical password system were highlighted. The countermeasures for each of the security threat were discussed. We believed this study could help the researchers who would like to do research on graphical password especially on recognition-based graphical password. In future, besides security aspects, we will focus on usability aspects research such as user login time and methods that can help users to recall their passwords.

Acknowledgement

This project was supported by the Postgraduate Re-search Grant (PPP) - PG005-2015B from the University of Malaya and also Fundamental Research Grant Scheme (FRGS) - FP071-2015A from the Ministry of Higher Education, Malaysia.

References

1. Por L. Y. and Lim X. T.: Issues, threats and future trend for GSP. in Proceedings of The 7th WSEAS International Conference on Applied Computer & Applied Computational Science, Hangzhou, China, 627–633 (2008)

2. Ho, P. F., Kam, Y. H. S., Wee, M. C., Chong, Y. N., Por, L. Y.: Preventing Shoulder-Surfing Attack with the Concept of Concealing the Password Objects' Information. The Scientific World Journal, (2014)
3. Por, L.Y., Ku, C.S., Islam, A., Ang, T.F.: Graphical password: Prevent shoulder-surfing at-tack using digraph substitution rules. The Frontiers of Computer Science, Accepted (2016)
4. Blonder, G. E.: ``Graphical Passwords'', United States Patent 5559961, Lucent Technologies, Inc. (Murray Hill, NJ), (1996)
5. Biddle, R., Chiasson, S., Van Oorschot, P. C.: Graphical passwords: Learning from the first twelve years. ACM Computing Surveys (CSUR), 44(4), 19 (2012)
6. Por, L. Y., Wong, K., Chee, K. O. : UniSpaCh: a text-based data hiding method using Unicode space characters. Journal of Systems and Software, 85(5), 1075–1082 (2012)
7. Por, L. Y., Delina, B.:Information hiding a new approach in text steganography. In Proceedings of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science, 2008, 689–695.
8. Por, L. Y., Delina, B., Ang, T. F., Ong, S. Y.: An enchanced mechanism for image steganog-raphy using sequential colour cycle algorithm. The International Arab Journal of Information Technology, 10(1), 51–60 (2013)
9. Por L. Y., Lai W. K., Alireza Z., Delina B.: StegCure: an amalgamation of different steganographic methods in GIF image. In Proceedings of the 12th WSEAS International Conference on Computers, Heraklion, Greece, 420–425 (2008)
10. De-Angeli, A., Coventry, L., Johnson, G., and Renaud, K.: Is a picture really worth a thou-sand words? Exploring the feasibility of graphical authentication systems. International Journal of Human-Computer Studies, 63, 128-152 (2005)
11. Passfaces^{TM}: The Science behind Passfaces, White paper. http://www.passfaces.com/enterprise/resources/white_papers.htm. Accessed 10 July 2017 (2000)
12. Brostoff, S., Sasse, M. A.: Are Passfaces more usable than passwords: a field trial investigation. In People and Computers XIV—Usability or Else! , 405-424: Springer London (2000)
13. Jansen, W., Gavrila, S., Korolev, V., Ayers, R., Swanstrom, R.: Picture password A visual login technique for mobile devices. (2003)
14. Davis, D., Monrose, F., Reiter, M. K.: On user choice in graphical password schemes. In USENIX Security Symposium, 13, 1-14 (2004)
15. Zhao, H., Li, X.: S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme. Proceedings of the International Conference on Advanced In-formation Networking and Applications Workshops, 2, 467-472 (2007)
16. De-Angeli A., Coutts M., Coventry L., Johnson G.: VIP: A Visual Approach to User Authentication. Proceedings of the Working Conference on Advance Visual Interfaces, 316-323 (2002)
17. Hayashi, E., Dhamija, R., Christin, N., Perrig, A.: Use Your Illusion: secure authentication usable anywhere. Proceedings of the 4th Symposium on Usable Privacy and Security, 35-45 (2008)

18. Gao, H., Liu, X., Wang, S., Liu, H., Dai, R.: Design and Analysis of a Graphical Pass-word Scheme. The 4th International Conference on Innovative Computing, Information and Control, 675-678 (2009)
19. Khot, R. A., Kumaraguru, P., Srinathan, K.: WYSWYE: shoulder surfing defense for recognition based graphical passwords. Paper presented at the Proceedings of the 24th Australian Computer-Human Interaction Conference, Melbourne, Australia (2012).
20. Perkovic, T., Cagalj, M., Rakic, N.: SSSL: Shoulder Surfing Safe Login. 17th International Conference on Software, Telecommunications & Computer Networks, 2009. SoftCOM (2009)
21. Por, L.Y.: Frequency of occurrence analysis attack and its countermeasure. The International Arab Journal of Information Technology, 10(2), 189-197 (2013)
22. Manjunath, G., Satheesh, K., Saranyadevi, C., Nithya, M.: Text-based shoulder-surfing resistant graphical password scheme. International Journal of Computer Science and Information Technologies, 5(2), 2277-2280 (2014)
23. Haque, M.A., Imam, B.: A New Graphical Password: Combination of Recall & Recognition Based Approach. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(2), 320-324 (2014)
24. Pooja K. S., Prajna V. D., Prathvi, Ashwini N.: Shoulder Surfing Resistance Using Graphical Password Authentication in ATM Systems. International Journal of Information Technology & Management Information System (IJITMIS), 6(1), 1-10 (2015)
25. Al-Ameen M. N., Wright M., and Scielzo S.: Towards making random passwords mem-orable: leveraging users' cognitive ability through multiple cues. CHI `15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp.2315-2324 (2015)
26. Assal, H., Imran, A., Chiasson, S.: An Exploration of Graphical Password Authentication for Children. https://arxiv.org/abs/1610.09743. Accessed Date: 16 June 2017 (2016)
27. Agrawal, S., Ansari, A. Z., Umar, M. S.: Multimedia graphical grid based text password authentication: For advanced users, 2016 Thirteenth International Conference on Wireless and Optical Communications Networks (WOCN), 1-5 (2016)
28. Por L. Y., Kiah M. L. M.: Shoulder surfing resistance using penup event and neighbouring connectivity manipulation. Malaysian Journal of Computer Science, 23(2), 121–140 (2010)

A Management Framework for Developing a Malware Eradication and Remediation System to Mitigate Cyberattacks

Nasim Aziz[1,2], Zahri Yunos[1], Rabiah Ahmad[2]

[1]CyberSecurity Malaysia, Malaysia
[2]Universiti Teknikal Malaysia Melaka, Malaysia
nasim_aaziz@yahoo.com, zahri@cybersecurity.my,
rabiah@utem.edu.my

Abstract. Malware threats are a persistent problem that interrupts the regular utilization of IT devices. For effective prevention of malware infections in computer system, development of a malware mitigation system needs to be developed. Malware mitigation system should encompass a thorough technical and management outlook to achieve an effective result. A Management Framework should thus be put in place to facilitate better management and effective outcomes of such a system. This research presents the identification, formulation and proposal of a Management Framework for the development of a malware eradication and remediation system to mitigate cyberattacks. The aim of this research is to construct a Management Framework that allows for the effective development of a malware eradication and remediation system. The method used in this work is qualitative research (observation and interviews) at organizations that have implemented similar systems. The framework covers specific areas that refer to the management of people, process and technology in designing a malware eradication and remediation system.

Keywords: Advanced Persistent Threat (APT), Critical National Information Infrastructure (CNII), Information Technology (IT), Internet of Things (IoT), Malicious Software (Malware)

1 Introduction

Information Technology (IT) instruments in this age of technological advancement improve the condition and efficiency of people livelihood. In this Internet of things (IoT) era, IT devices are internetworked and created specifically in order to make human activities easier and more efficient. Among the possible outcomes of IT technology are self-driven vehicles, which are being developed by automotive manufacturers. However, in this technology time, possible cyber threats such as malicious software (malware) are persistently being created to perform specific attacks on technological devices. In the example of self-driven vehicles, vehicles infected by malware may not to be driven to the right destination to which they were programmed, or

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_50

passengers may be kept for ransom in vehicles by malware attackers. Malware attacks using stealthy continuous hacking methods exploit vulnerabilities in computer systems with the goal of attaining financial benefits, political advantages, IT infrastructural control, or even simply for information gathering to take advantage of victims [1].

To curb such malware threats, an advanced technological mitigation system that can eradicate harmful software needs to be developed to contain successful malware infections. Developing a malware eradication and remediation system to mitigate cyberattacks is therefore crucial to maintaining the safety and security of using IT devices and systems. Through a malware mitigation system, malware that infects a computer system can be detected and eradicated whilst allowing the computer system to be remediated as per the primary program intended. However, in developing such a malware mitigation system, the proper management of components, particularly People, Process and Technology as depicted in Fig. 1 needs to be considered to achieve the intended goal of mitigating cyberattacks.

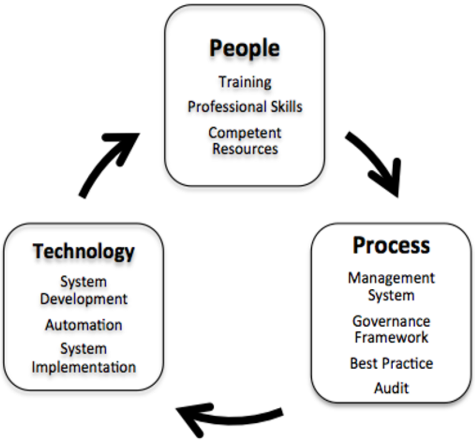


Fig. 1. Components of the Management Framework for Developing a Malware Eradication and Remediation System

Disruption to system infrastructure, and theft of financial data and intellectual property will undoubtedly drive investment away from countries whose computer systems are seen to be insecure [2]. As such, the requirement to manage the process of developing a malware eradication remediation system for mitigating cyberattacks on organizations' infrastructure is crucial for ensuring newly developing countries are safe and secure against these types of attack. A management framework is thus required in order to identify the best solution for developing an effective malware eradication system.

2 Related Works

The development and implementation of a malware eradication and remediation system can improve cyber security against malware attacks. It is crucial to safeguard data and information, especially concerning national interest being kept safe and sound from exploitation and theft. The development of such mitigation system will help research activities and improve the detection of new malware and cyber threats from damaging a country's Critical National Information Infrastructure (CNII) as well as public IT devices for the benefit of public safety and security. Nevertheless, it is important to establish a Management Framework for the development of a malware eradication and remediation system. A thorough and intensive examination of the types of malware and mitigation systems should be conducted in order to achieve the best solution. A literature review of Management Frameworks for developing such malware mitigation system is therefore required in order to gain the best understanding and solutions based on an extensive theoretical review.

In order to develop any infrastructure, it is imperative to have a Framework Design as to establish the core basis of what the intended outcome to be accomplished is. A framework is defined as the relevant objects and coherence scaffold of aspects that must be considered during the design and implementation process [3]. Framework design can be considered to be an extensive process of various components. Under the National Institute of Standards and Technology, US Department of Commerce [4], it summarizes a guide for applying a Risk Management Framework to its federal information system that includes components such as Categorizing, Selecting, Implementing, Assessing, Authorizing and Monitoring. Researchers and organizations have proposed many framework designs, but what is important is that the purpose and function of developing the framework is achieved.

Malicious software (Malware) refers to any type of malicious software used in attempts to infect technological systems, be they computers, phones or tablets. Viruses are specific types of malware that perform specific types of attacks that are basically designed to replicate or spread to computer systems, while malware consists of a wide range of malicious codes, such as viruses, spyware, adware, nagware, trojans, worms and other malicious software [5]. Hackers use malware for a variety of purposes, most of which involve extracting personal information, stealing money or intellectual property, or preventing owners from accessing their devices. Over the Internet, coordinated malware attacks are considered the most dangerous threats to system security [6]. Historically, viruses and worms have been respectively created just for fun and to repair computer systems. Malware came into existence in the 1980s from a virus named Brain that "booted up" personal computers after users inserted a floppy disc, while the Morris worm came to attention when more than 6000 computers were affected by its activity [7].

Types of malware symptoms or activities vary among the intended malware to be used. However, a similar indication of malware compromise is based on the activity of an infected computer system. Symptoms of an infected computer system can be an increase in the computer's processing usage, a decrease in computer usage or web browser capability, difficulty connecting to networks, and computer system freezing or crashing. Other serious signs of a malware attack can also be alteration or deletion of file information, presence of unfamiliar files, programs or desktop icons. Automatic running of programs, turning on/off of IT devices, or reconfiguring of programs and abnormal device system behaviour such as the lock down of computers' access control are all activities of malware in a computer system.

Similar to other countries, the Malaysian Internet landscape encounters cyberattacks due to the widespread use of IT devices connected to the Internet. According to an Internet Users Survey in 2016 [8], it was revealed that in 2015 two-thirds to three-fourths of Malaysia's population were part of the online community. This shows that potentially, malware infections of IT devices used by Malaysians are high, as the Internet is viable to be acquired easily in the country. According to CyberSecurity Malaysia statistics [9] and based on feeds to the Cyber999 help centre, 2016 witnessed 2,026,276.00 Malaysia Botnet Drones by unique IPs and 1,130,056.00 Malware Infections by Unique IPs. Figure 2 shows yearly statistics since 2011 till 2016) on malware feeds to CyberSecurity Malaysia.

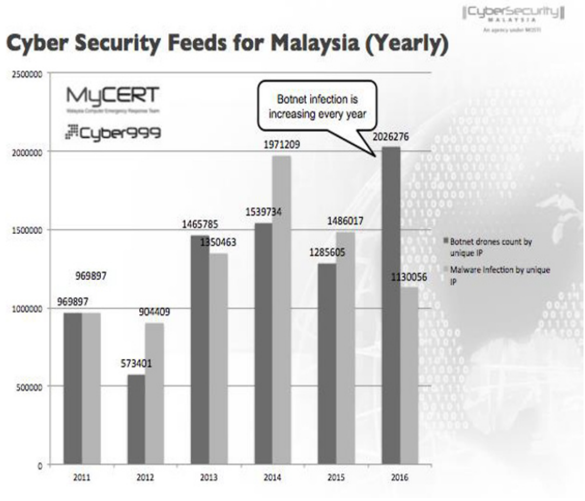


Fig. 2. Malaysia Botnet Drones and Malware Infections (2011-2016)

In May 2017, the WannaCry ransomware created worldwide panic as the malware targeted computers running the Microsoft Windows operating system. Malaysia was lucky not to have been greatly affected by the attack, although two organizations were hit [10]. However, looking at the situation caused by the attack, the Malaysian Internet landscape also experienced alarm bells due to the uncertainty of cyberattacks.

According to a survey conducted by Quann[11] in 2017, most companies in Malaysia are not ready for cyberattacks due to a lack of security preparedness, the absence of a Security Operation Centre (SOC) and no proper IT security awareness program for its employees.

Management framework has not been standardized to combat malware threats. Various organizations and researchers around the world have come up with relevant antiviruses, mitigation systems and software in order to eradicate malware. Researchers from the State University of New York and University of California proposed a system called Malware-Aware Processors (MAP) that was designed to improve ways of detecting online malware [12]. The system was introduced and applied, whereby the always-on nature of MAP prioritizes the scanning order of processes. It allows for the most anomalous software processing to be scanned first, thus making it difficult for malware to avoid detection. According to Almarri and Sant [6], a framework needs to be created to detect and analyse malware through the ability to perform both as a detector as well as a warning system. The proposed framework comprises three phases: Phase 1: Malware Acquisition Function, Phase 2: Detection and Analysis, and Phase 3: Database Operational Function.

In 2014, a service-oriented malware detection framework called 'SmartMal' was introduced. It was claimed to be the first to combine service-oriented architecture (SOA) concepts with state-of-the-art behaviour-based malware detection methodologies. A research paper [13] introduced the SOA concept for malware detection mechanisms in order to construct a distributed malware detection framework with a behaviour analysis model.

Banescu [14] designed FEEBO, a framework to conduct empirical experiments on the effects of behaviour obfuscation on malware detection. It is used to apply certain obfuscation transformations to the externally visible behaviour of malware samples. The empirical evaluation framework for obfuscating known malware binaries was first assessed in the article, after which the impact on the detection effectiveness of different n-gram based detection approaches was investigated. In order to curb on-going online threats that are ready to attack computer systems, a management framework is therefore necessity to guide and inform necessary individuals or organisations of how to react in case of malware attacks and ultimately manage such cyber threats.

This conceptual research paper is written to address the need for development of management framework that is effective in mitigating malware in computer system. It is based upon a malware mitigation system that will be developed later in comparison with existing framework designed from other existing mitigation systems as per example above (section D). This is to establish of a documented paper in developing a malware mitigating system, giving focused on the management framework that is of academic quality.

3 Management Framework to Mitigate Cyberattacks

The underground economy gives rise to creation of malware attacks that are profit-oriented and thievery of critical information. Specific malware attacks on critical information of the government as well as monetary institutions along with related entities are a concern to the government and security field. According to the Australian Cyber Security Centre [15], cyber adversaries are aggressive and persistent in their efforts to compromise Australian networks and information.

Mitigation systems to engage and deter successful malware attacks are therefore required by organizations that appreciate a secure IT environment. Government institutions, financial entities, military forces and other organizations that are well-equipped with IT infrastructure thus require a well-developed malware mitigation system that will prevent malware infections in its computer system. Nevertheless, information regarding malware eradication and remediation systems is limited and inadequate to assist researchers to improve the development of such a system.

A Management Framework to mitigate cyberattacks by understanding and repairing damaged systems is a necessity to the development of a malware eradication and remediation system. However, developing and implementing a prototype system is something that needs to be attempted, as it is of significant value to any government or organization. This study thus emphasizes three key tenets: People, Process and Technology.

3.1 Aim and Objectives

Due the ineffectiveness of current practise in preventing infection of malware in computer systems when it involves multi-organization, it is the aim of this research to develop a Management Framework model for the development of a malware eradication remediation system to mitigate cyberattacks. The objective of this research is therefore to answer the following research questions:

1. Q1. How can the effectiveness of a malware eradication and remediation system in repelling successful malware attacks on a computer system be measured?
2. Q2. What is an effective and suitable Management Framework model to act as a governance structure of a malware eradication and remediation system?

Through this research, an attempt is made to understand the types of malware that exist and what management models can be considered in developing a malware mitigation system. Based on the problem in Q1, a framework will be facilitated that addresses the management of three components (people, process, technology) of a malware mitigation system and provides solutions for mitigating malware attacks particularly in the Malaysian landscape.

3.2 Methodology

The qualitative research method is proposed for the current work. The design of this research takes into consideration of the Descriptive Design and using the Qualitative Research Methodology to assists in identifying problems existing in the environment, in an attempt to find the best solution to improve the situation. Descriptive design describes a phenomenon, current situation or characteristics of a group of organization, people, etc. The objective of descriptive research is to describe things, such as the IT security situation, its potential, and acceptance for a new concept. The need for a management framework to develop a malware eradication and remediation system at any organization is analyzed in this research.

Qualitative research methodology meanwhile is based on interpretations from information attained through research study that can help identify the best solution derived from the data collected. Through qualitative research method, the aim is to 'understand' the contributing factors that allow for a situation to happen [16]. This method also provides 'explanation' of occurrences acquired through observation and/or reviews, 'rationalization' of circumstances focused on small-scaled focus groups through interviews and observation, and 'recognition' of information attained from researchers who become participants themselves.

The research hypothesis is that currently, most organizations use some Anti-Virus, IDS, BDS and firewall to defend themselves against malware attacks. If an organization establishes a management framework to guide the development of a malware eradication remediation system, it could help the organization manage malware attacks effectively. The hypotheses of the research are as follows:

1. A malware eradication and remediation system is a necessity of any IT-based organization.
2. A reliable management framework will allow an organization to develop a malware eradication and remediation system that can effectively mitigate malware attacks.

Data sources are identified from analysing various organizations that have developed means of managing malware intrusions or attacks. Data is collected from primary and secondary sources. Primary sources involve observation, interviews, meetups with professional security practitioners and surveying organizations that have already developed malware mitigation systems. Secondary resources include research materials from other organizations and previously published research papers. Majority of research studies encounter certain limitations that might affect the overall research outcome. The current study limitation assumptions are as follows:

1. Assessing the effectiveness of malware mitigation system due to obtaining confidential documents and resources disclosed by the parties involved are limited.
2. Availability of reference material and information relating to management framework for comparing the ideal system in terms of malware eradication and remediation are scarce.

Initial result shows that malware infection is increasing in Malaysia and it is significant to develop a system to mitigate this type of cyberattack. It is also important to establish a management framework for the mitigation system in order to achieve an effective result. Based on the data collected through primary and secondary sources, the result of the research should determine an effective Management Framework for developing a malware mitigation system to eradicate and remediate malware outbreaks and prevent cyberattacks that involves People, Process and Technology.

4 Conclusion

Findings from this conceptual paper can solve in mitigating cyberattacks by improving the current practice in detecting, eradicating and remediating computer system that have been infected by malware. The outcome of the research will allow formulating of a Management Framework to mitigate malware attack that is inclusive of organization components (People, Process & Technology). The main contribution of this research is the formulation of an effective Management Framework model for the development of a malware eradication and remediation system. This will be accomplished by determining the factors that will facilitate the successfulness of managing the system development. This research will also assist other developers to design and develop a malware mitigation system that is effective in eradicating and remediating malware infections in computer systems at their own organizations. Upon identifying and documenting the framework, it is believed that further enhancements of mitigating cyberattacks particularly from malware can be reduced.

Acknowledgements

We thank CyberSecurity Malaysia and its management for the support in conducting this research study. Appreciation also goes to Universiti Teknikal Malaysia Melaka (UTeM) for their guidance in performing the research. The authors encourage reviews from fellow researchers and IT security practitioners for insightful comments and suggestions

References

1. E. M. Rudd, A. Rozsa, M. Günther, and T. E. Boult, A Survey of Stealth Malware: Attacks, Mitigation Measures, and Steps Toward Autonomous Open World Solutions, (2016) pp. 1–28,.
2. D. A. Wahab, Facing cyberattacks in 2016 and beyond, *http://www.thestar.com.my/tech/tech-opinion/2016/01/28/facing-cyber-attacks-in-2016-and-beyond/*, (2016).
3. J. D. , K.Grant, Leading Issues in Knowledge Management. Academic Conference and Publishing Limited., (2015).
4. NIST, Framework for Improving Critical Infrastructure Cybersecurity," *Natl. Inst. S*, (2014) pp. 1–41,.
5. A. Henry, The Difference Between Antivirus and Anti-Malware, *https://lifehacker.com/the-difference-between-antivirus-and-anti-malware-and-1176942277*, (2013).

6. S. Almarri and D. P. Sant, Optimised Malware Detection in Digital Forensics, *Int. J. Netw. Secur. Its Appl.*, vol. 6, no. 1, (2014), pp. 1–15,.
7. A. Kaushik, *Sailing Safe in Cyberspace: Protect Your Identity and Data*. SAGE Publications India, (2013),.
8. Malaysian Communications and Multimedia Commission, Internet Users Survey 2016, (2016).
9. MyCERT, MyCERT Incident Statistics,"*https://www.mycert.org.my/statistics/2017.php*, (2017) .
10. R. S. Bedi and A. Chan, WannaCry strikes two Malaysian companies, *http://www.thestar.com.my/news/nation/2017/05/16/wannacry-strikes-two-msian-companies-expert-first-organisation-infected-last-saturday/*, (2017) .
11. Quann, Malaysian companies are unprepared for cyber attacks , a survey by Quann reveals, *https://www.quannsecurity.com/downloads/resource/MYIDC.pdf*, (2017). .
12. M. Ozsoy, C. Donovick, I. Gorelik, N. Abu-Ghazaleh, and D. Ponomarev, Malware-aware processors: A framework for efficient online malware detection, *2015 IEEE 21st Int. Symp. High Perform. Comput. Archit. HPCA 2015*, (2015) pp. 651–661,.
13. C. Wang, Z. Wu, A. Wang, X. Li, F. Yang, and X. Zhou, SmartMal: A service-oriented behavioral malware detection framework for smartphones, *Proc. - 2013 IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2013 2013 IEEE Int. Conf. Embed. Ubiquitous Comput. EUC 2013*, (2014) vol. 2014, pp. 329–336,.
14. S. Banescu, T. Wüchner, M. Guggenmos, M. Ochoa, and A. Pretschner, FEEBO: An Empirical Evaluation Framework for Malware Behavior Obfuscation, *arXiv Prepr. arXiv1502.03245*, (2015).
15. ACSC, 2015 Threat Report, *Aust. Cyber Secur. Cent. Threat Rep. 2015*, (2015) p. 29,.
16. Z. Yunos, N. Mohd, A. Ariffin, and R. Ahmad, Understanding Cyber Terrorism From Motivational Perspectives: A Qualitative Data Analysis, in *16th European Conference on Cyber Warfare and Security*, (2017).

A Review on Energy Efficient Path Planning Algorithms for Unmanned Air Vehicles

Sanjoy Kumar Debnath, Rosli Omar, Nor Badariyah Abdul Latip

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia,
Parit Raja, Batu Pahat, Johor-86400, Malaysia
sanjoyyy@yahoo.com, roslio@uthm.edu.my, norbadariyah92@gmail.com

Abstract. Unmanned Aerial Vehicle (UAV) is a type of autonomous vehicle for which energy efficient path planning is a crucial issue. The use of UAV has been increased to replace humans in performing risky missions at adversarial environments and thus, the requirement of path planning with efficient energy consumption is necessary. This study analyses all the available path planning algorithms in terms of energy efficiency for a UAV. At the same time, the consideration is also given to the computation time, path length and completeness because UAV must compute a stealthy and minimal path length to save energy. Its range is limited and hence, time spent over a surveyed territory should be minimal, which in turn makes path length always a factor in any algorithm. Also the path must have a realistic trajectory and should be feasible for the UAV.

Keywords: Energy efficient, UAV, Path planning, Optimal path.

1 Introduction

The use of UAVs has been increased to perform missions, such as weather forecasting, traffic control and rescue people [1]. The mission may be in a cluttered and obstacle-rich environment; for example in an urban area and hence, it is important for a UAV to adopt a path planning algorithm ensuring the traversed path to be collision-free and optimal in terms of path length. However, optimal path only is not enough as it may cause the UAV to consume more energy than a suboptimal one. Most common problem of UAV path planning is to fly from a given starting point to a target point through a set of obstacles [2]. These obstacles may not be fixed at one location and can pop up during the fly. An energy efficient path planning must ensure that the method/algorithm can create a safe and optimal path and, simultaneously can minimize the travel duration and save energy/fuel. This paper discusses different approaches of path planning which also considers the energy consumption and path length.

The configuration space (C-space), which is a most commonly used technique for path planning, provides detailed position information of all points in the system and is the space for all configurations. It assumes that the UAV as a point and adds the area of the obstacles so that the path planning can be done more efficiently. C-space is obtained by adding the UAV radius while sliding it along the edge of the obstacles

The original version of this chapter was revised: A correction to this chapter can be found at https://doi.org/10.1007/978-981-13-2622-6_56

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_51

and the border of the search space. An illustration of a C-space for a circular UAV is shown in Fig. 1.

In Fig.1(a), the obstacle-free area is represented by the white background while the solid dark area represents the obstacles' region. The UAV is denoted by a black dot circled with gray color and three pre-planned paths are represented by dotted, semi-dotted and solid lines to reach the target/goal configuration Q_{goal} from start/initial configuration Q_{init} considering that the C-space is not created. Conversely, when the workspace is considered as C-space, as shown in Fig. 1(b), the UAV has only one feasible path. This also reveals that the free space Q_{free} has been reduced while the obstacles' region Q_{obs} has been increased. Therefore, C-space denotes the real free space area for the movement of UAV and ensures that the vehicle or UAV must not collide with the obstacle. The popularity of C-space method in path planning is due to its use of uniform framework to compare and evaluate various algorithms [4]. The UAV's or robot's path planning can be classified in three ways namely combinatorial, sampling based and biologically inspired methods as illustrated in Fig. 2.

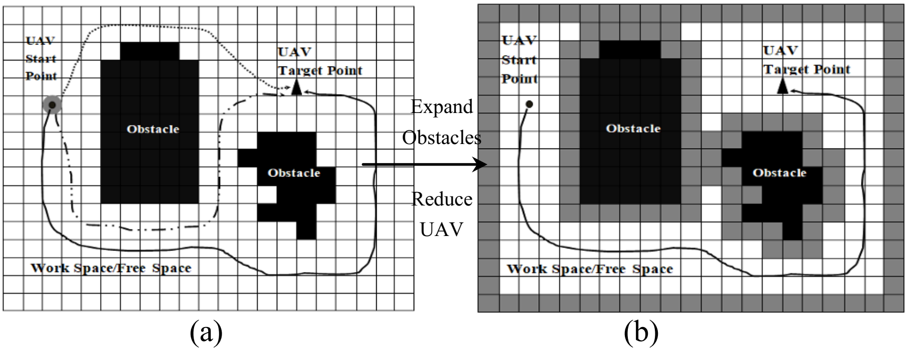


Fig. 1. Configuration space for a UAV path planning

2 Combinatorial Path Planning

Combinatorial method applies C-space concept to the workspace representation methods such as cell decomposition (CD), potential field (PF), visibility graph (VG) and Voronoi diagram (VD), to name a few, coupled with graph search algorithms like Dijkstra's, A-star, Breadth First Search and Depth First Search so that a collision-free energy efficient path can be found [5]. Researchers already proposed several techniques on path planning classified as roadmap method such as VG and VD, and other methods like CD and PF. The C-space representation allows efficient path planning techniques based on roadmap and cell decomposition to obtain a solution. The roadmap captures the connectivity within Q_{free} using a graph or network of paths. In a roadmap, nodes are considered as points in Q_{free} and two nodes are adjoined by an edge that must be within Q_{free}. A set of collision-free paths from an initial configuration Q_{init} to a goal configuration Q_{goal} builds the roadmap that uses several steps for path planning. Firstly, it connects the nodes with edges in free C-space area to build a network of graph. After that Q_{init} and Q_{goal} are associated with the network to conclude the roadmap. A series of line segments constructs a collision-free optimal path that can be explored within Q_{free}.

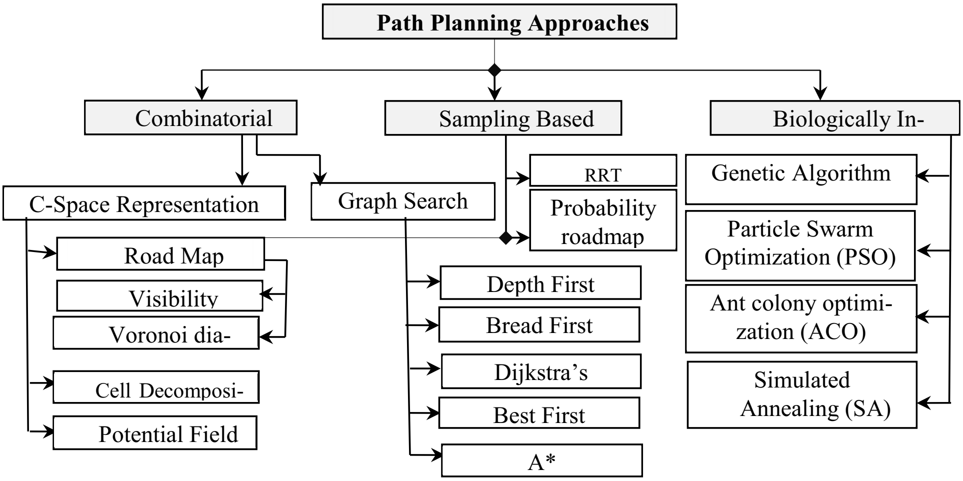


Fig. 2. Classification of path planning algorithms

Visibility Graph (VG). It is a popular and efficient path planning method under roadmap category where the available paths consist of waypoints that are also the nodes of obstacles and this makes the paths semi-collision-free. Vertices V of a VG graph comprise Qinit, Qgoal and the polygonal obstacles vertices [35]. The edges are made of the edges of obstacles and edges joining all pairs of vertices that lie in the Qfree. Lozano-Perez and Wesley initially proposed VG for path planning within an environment consisting polyhedral obstacles [7]. Oomen et al. used VG method in [9] to resolve the path planning of an autonomous mobile robot in an unexplored obstacle-filled environment. Research on the reduction of VG's complexity was projected in [2]. Both approaches claimed to be suitable and energy efficient for path planning in real-time. VG based algorithm, known as Equilateral Space Oriented Visibility Graph (ESOVG), was proposed to reduce the number of obstacles for a car-like robot during its path planning [11] and is capable of saving energy and finding optimal path with less computation time.

Voronoi Diagram (VD) - A set of regions built by dividing up the C-space makes the VD. Each region and all points in it correspond to one of the sites [12]. It generates Voronoi edge that is equidistant from all the points of the obstacles' area in C-space. Hence, it cannot generate optimal paths. A dynamic path planning for multi-robot was proposed to cover the sensor-based narrow environments assuming the energy capacities of the mobile robots based on generalised VD (GVD) graph [14]. A path planning method for Unmanned Surface Vehicle (USV), which operates at sea integrates the VD, VG and Dijkstra search algorithm and is able to find a collision-free and energy efficient path where up to 21 % energy can be saved [15].

Cell Decomposition (CD) – At outdoor environment it is a popular path planning method where the workspace is decomposed into discrete, non-overlapping, rectangular or polygonal shaped cells in between start and target points producing a continuous path [16] or connectivity graph. Here, the obstacle-free cells must be completely free from any obstacle or its part; else they are identified as occupied. There are several variants of CD including Regular Grid, Adaptive Cell Decomposition and Exact Cell Decomposition. Unfortunately, CD has several drawbacks like generation of infeasible solutions, combinatorial explosion and limited granularity. A cell decomposition approach for trade-off safe and short path was considered in [18] by choosing the weight values. Another cell decomposition method efficiently covers the optimised path over cells as per the distance between the centroids of cells and reduces the rate of energy consumption and operational time [19].

Potential Fields (PF) - In PF, Qgoal and obstacles have attractive and repulsive potentials respectively. The goal configuration and the obstacles produce a potential field at which the robot travels. PF was firstly suggested by Khatib [6], which considered a UAV as a point under the influence of the fields produced by Q_{init}, Q_{goal} and obstacles within C-space. The resultant force of the field on the robot determines the vehicle motion direction. As potential field method directs the vehicle towards a minimum in the field, it is not guaranteed that the minimum is the global minimum. A global off-line path planning approach is implemented using an energy-based approach known as Artificial Potential Field (APF) for Multi-Robot Systems (MRSs). Based on the potential field, an improved artificial potential field (APF) UAV path planning technique was introduced and it is more effective in finding the shortest path [21]. Another potential field technique uses the kinematics of a six wheel rover for motion on rough 3D terrain where relative significance of the paths is obtained from four different cost functions with respect to energy, traction force, slip and deviation from a straight line. Extensive experiments and simulations proved that this method is better in obtaining paths [22].

3 Sampling-Based Path Planning

Sampling-based motion planning methods are used during the search within configuration space when information is obtained from a collision detector. The geometric model, a sequence of sampling-based algorithm, depends on potential configuration and checks collision so that the validity of the configuration can be verified and it ends with the matching of a configuration with the goal configuration. Since the collision checking is done as required, thus this algorithm lacks the knowledge about the presence of the object in the configuration space. There are two popular methods in sampling based path planning, namely Rapidly–exploring Random Tree (RRT) and Probabilistic Roadmap (PRM) which are elaborated below [5].

Rapidly-exploring Random Tree (RRT). This algorithm efficiently searches in increment without the outline of high-dimensional spaces by constructing a space-filling tree from randomly drawn samples within the workspace. It is fundamentally influenced to grow towards the large areas of problems that are not searched. This was proposed by LaValle and Kuffner Jr in [24, 25] as an easy solution to handle problems with obstacles and differential constraints for autonomous robotic motion planning. The computation time increases depending on the size of the generated tree. The resulting path from RRT is always not optimal. But it is quite easy to find a path for a vehicle with dynamic and physical constraints, and produces minimum number of edges. A RRT based path planning algorithm generates a cost-efficient path to satisfy the requirements of the mission stated in linear temporal logic (LTL) [26] where the cost function comprises the hazard levels, the energy consumption and the wireless connectivity. Kamarry et al. used a compact RRT illustration by decreasing the redundancy of the nodes and the number of discarded samples. The processing time of the tree growth and the computation cost were also reduced for which the path length was shortened and the energy was saved [27].

Probabilistic Roadmap (PRM). It is a motion planning algorithm that takes random samples from the configuration space by checking the available free space and avoiding the collisions to determine a path. A local planner is used to join these configurations with nearby configurations. A graph search algorithm is applied after adding the initial and goal configurations to determine a path. It has two phases, i.e., construction and query phase. Approximating the motions, a roadmap (graph) is built in the construction phase. Whereas, the start and goal configurations are connected to the graph in the query phase, and then the path is produced by a graph search algorithm. The obtained path often has poor quality and as a result of randomness, it represents the free space connectivity. This method may also be incomplete, i.e., it is unable to figure out a path between two locations although the path exists connecting these locations in the presence of narrow passage. Moreover, it is tough to know about any existing path unless it is found [28, 29]. On the other hand, PRM is probabilistically optimal and complete with reasonable computation time. Chung et al. [30] presented a PRM based method for low-cost path planning of a UAV in a spatially varying wind field using a biased sampling-based path search technique. They minimised its energy consumption by finding time efficient path planning through the available flight region.

4 Biological-Based Path Planning

Biologically inspired method is based on biology, computer science, mathematics and artificial intelligence, mainly on machine learning, and is a major subset of natural computation. It can also be stated as the combination of connectionism, social behavior and emergence. In this technique, the living phenomenon is modelled by using computers and concurrently it tries to make the improved use of computer for a better life [31]. There are numerous methods for an energy efficient path planning of a UAV among which a few are discussed below:

Genetic Algorithm (GA). Constrained and unconstrained, both of these optimization problems can be resolved by using the natural selection process of driving biological evolution and it continuously changes a population of individual results [32]. But it cannot guarantee any optimal path. Local minima might occur in narrow environments and thus, it gives less safety and narrow corridor problem. GA is computationally expensive and practically not complete. Li et al. utilised GA for traversing the energy map in outdoor environment with 3D to avoid local optimisation. The model was incorporated with an estimated energy consumption formula that was served as an input of the energy consumption map and as a result the energy-optimal paths were obtained by GA [33]. Dogru et al. optimised the Coverage Path Planning (CPP) using GA in terms of energy consumption by considering the limitations of natural terrains such as obstacles and relief. This method seems to be effective for a mobile robot performing CPP as per the simulation result to reduce the energy consumption [34].

Particle Swarm Optimization (PSO). This is a classical meta-heuristic population-based algorithm, originally introduced by Kenney and Eberhart in 1995, to resolve global optimization issues based on swarming or collaborative behavior of biological populations. A PSO based path planning algorithm tested the problems of multi-objective path planning models verified with the focus on robot's energy consumption and path's safety [17]. Another PSO based method, known as improved particle swarm optimization (IPSO) with an improved gravitational search algorithm (IGSA), was proposed to reduce the maximum length of the path which in turn should minimize the travel time for all robots to reach their respective destination within the environment along with the optimization of the energy with respect to the turn’s number and arrival time [3].

Ant Colony Optimisation (ACO). This is meta-heuristic and probabilistic technique, developed by Dorigo in 1992 [23] based on the behavior of the ants to search their food and create the paths after locating its source. However, it suffers inherent parallelism. At the same time, positive feedback stimulates the quick discovery for good results. Lee et al. proposed an energy-based ant colony optimisation algorithm for battery-powered electric automobiles to attain the optimised energy-conserving path [10]. Zaza et al. presented an improved Ant ACO to resolve various Vehicle Routing Problems (VRPs) by utilising the task allocation and route planning methods for a UAV. This new algorithm can be used for collision avoidance penalties and it can change the travelling time of each task [8].

Simulated Annealing (SA). This is a probabilistic meta-heuristic technique to approximate the global optimum of a given function within a discrete space for a large search. It is preferable for the alternatives, such as gradient descent, where it is problematic and as well important to find an approximate global optimum than obtaining a precise local optimum within a given time. The enhanced SA is capable of giving a near-optimal or optimal path solution for various dynamic workspaces with less processing time and can also improve the real-time computational efficiency [13]. Turker et al. solved the multiple UAVs' path planning issues using parallel SA algorithms and it was executed on parallel computers [20].

5 Discussion

The information in previous section reveals that VG is more energy efficient than VD in combinatorial method under roadmap technique. It is necessary for CD to adjust with the situation as required; e.g., in exact CD, the cells are not predefined, but they are selected based on the location and shape of the obstacles within the C-space [16].

Table 1 tabulates all the path planning methods versus their properties in terms of path optimality, computation time, real time capability, memory usage and completeness. Sometime PF could not find the goal because of local minima issue. In sampling based method, RRT does not always provide optimal result and PRM is expensive without any guarantee of finding the path. GA is currently used for energy efficient path planning but it also cannot guarantee to produce optimal path because local minima may occur in narrow environments. Moreover, GA is computationally expensive and practically not complete. PSO has real time effect but it can easily falls into local optima in many optimization problems. Furthermore, there is no general convergence theory applicable to PSO in practice and for multidimensional problems, its convergence time is also uncertain. ACO does a blind search and thus, is not optimistic and suitable for energy saving path planning. Apart from very slow and very high cost functions, SA is also not capable of finding optimal path.

6 Conclusion

Researchers used a number of path planning algorithms. Efficient path planning algorithm (i) can find an optimal collision-free path, (ii) is complete, (iii) has minimal computation time, and (iv) produces energy efficient path. Currently path planning approach is multi-dimensional. This paper focuses on the classification of path planning algorithms that consider energy efficiency and their nature of motion, advantages and drawback. Based on the objective of the UAV's mission and considering the outcome of the available path planning algorithms, such as computation time, completeness and safety [16], they can be optimised to produce the energy efficient path planning algorithm.

Acknowledgments. This work was supported by UTHM and funded by GPPS and Fundamental Research Grant Scheme (FRGS) with vot numbers of U457 and 1489 respectively.

Table 1. Comparison of Different Path Planning Methods Properties

C-Space Representation Techniques		COMBINATORIAL METHOD			
		Method	Optimal Path	Computation Time	Completeness
	Road map	Visibility Graph	√	×	√
		Voronoi Diagram	×	√	√
	CD	Regular Grid	×	×	×
		Adaptive Cell Decomposition	×	√	√
		Exact Cell Decomposition	×	×	√
	PF	Potential Field	×	√	×
SAMPLING BASED METHOD					
RRT			×	×	√
PRM			×	√	×
BIOLOGICALLY INSPIRED METHOD					
GA			×	√	×
PSO			×	×	√
ACO			×	×	√
SA			×	√	√

References

1. Liu Z, Sengupta R. An energy-based flight planning system for unmanned traffic management. InSystems Conference (SysCon), 2017 Annual IEEE International (2017) Apr 24 (pp. 1-7). IEEE.
2. R. Omar and D-W Gu., Visibility line based methods for UAV path planning. In Proceedings of the International Conference on Control, Automation and Systems (ICCAS-SICE), (2009) August 18, pp. 3176-3181.
3. Das PK, Behera HS, Panigrahi BK. A hybridization of an improved particle swarm optimization and gravitational search algorithm for multi-robot path planning. Swarm and Evolutionary Computation. 28, 14-28 (2016)
4. Dadi Y, Lei Z, Rong R, Xiaofeng X. A new evolutionary algorithm for the shortest path planning on curved surface. InComputer-Aided Industrial Design and Conceptual Design, (2006). CAIDCD'06. 7th International Conference on 2006 Nov 17; IEEE.
5. LaValle, Steven M. Planning algorithms. Cambridge university press, (2006).
6. Khatib, Oussama. "Real-time obstacle avoidance for manipulators and mobile robots." The international journal of robotics research 5.1 (1986): 90-98.
7. Lozano-Pérez T, Wesley MA. An algorithm for planning collision-free paths among polyhedral obstacles. Communications of the ACM. (1979), 22(10), pp.560-70.

8. Zaza T, Richards A. Ant Colony Optimization for routing and tasking problems for teams of UAVs. In Control (CONTROL), 2014 UKACC International Conference, (2014) Jul 9; IEEE.
9. Oommen, B., S. Iyengar, N. N. S. V. Rao, and R. Kashyap. "Robot navigation in unknown terrains using learned visibility graphs. Part I: The disjoint convex obstacle case." IEEE Journal on Robotics and Automation. 3, 6 (1987).
10. Lee KT, Huang SH, Sun SH, Leu YG. Realization of an Energy-Based Ant Colony Optimization Algorithm for Path Planning. In ICSSE (2015), pp. 193-199.
11. Latip, Nor Badariyah Abdul, Rosli Omar, and Sanjoy Kumar Debnath. "Optimal Path Planning using Equilateral Spaces Oriented Visibility Graph Method." International Journal of Electrical and Computer Engineering (IJECE) 7, 6 (2017).
12. S. Fortune, Voronoi Diagrams and Delaunay Triangulations, in: D.A. Du and F.K. Hwang, Editor, Euclidean Geometry and Computers, World Scientific Publishing, Singapore, (1992).
13. Miao H, Tian YC. Dynamic robot path planning using an enhanced simulated annealing approach. Applied Mathematics and Computation. 222,420-37 (2013).
14. Yazici, Ahmet, et al. "A dynamic path planning approach for multirobot sensor-based coverage considering energy constraints." IEEE transactions on cybernetics 44.3 (2014): 305-314.
15. Niu H, Lu Y, Savvaris A, Tsourdos A. Efficient Path Planning Algorithms for Unmanned Surface Vehicle. (2016) Dec 31; IFAC-Papers OnLine.
16. J. Giesbrecht and Defence R&D Canada. Path planning for unmanned ground vehicles. Technical Memorandum DRDC Suffield TM 2004-272, (2004).
17. Davoodi M, Panahi F, Mohades A, Hashemi SN. Clear and smooth path planning. Applied Soft Computing. 32:568-79 (2015).
18. Abbadi A, Přenosil V. Safe path planning using cell decomposition approximation. Distance Learning, Simulation and Communication. (2015) May 19; 8.
19. Janchiv A, Batsaikhan D, hwan Kim G, Lee SG. Complete coverage path planning for multi-robots based on. InControl, Automation and Systems (ICCAS), 11th International Conference on (2011) Oct 26; IEEE.
20. Turker T, Yilmaz G, Sahingoz OK. GPU-Accelerated Flight Route Planning for Multi-UAV Systems Using Simulated Annealing. InInternational Conference on Artificial Intelligence: Methodology, Systems, and Applications (2016) Sep 7; Springer International Publishing.
21. Chen, Yong-bo, et al. "UAV path planning using artificial potential field method updated by optimal control theory." International Journal of Systems Science 47, 6 (2016).
22. Raja, Rekha, Ashish Dutta, and K. S. Venkatesh. "New potential field method for rough terrain path planning using genetic algorithm for a 6-wheel rover." Robotics and Autonomous Systems 72 (2015).
23. Dorigo, Marco. "Optimization, learning and natural algorithms." Ph. D. Thesis, Politecnico di Milano, Italy (1992).
24. LaValle, Steven M. "Rapidly-exploring random trees: A new tool for path planning." (1998).
25. LaValle, LaValle SM, Kuffner Jr JJ. Randomized kinodynamic planning. The International Journal of Robotics Research. 20,5 (2001)
26. Cho, Kyunghoon, et al. "Cost-Aware Path Planning Under Co-Safe Temporal Logic Specifications." IEEE Robotics and Automation Letters 2.4 (2017).
27. Kamarry S, Molina L, Carvalho EÁ, Freire EO. Compact RRT: A New Approach for Guided Sampling Applied to Environment Representation and Path Planning in Mobile

Robotics. InRobotics Symposium (LARS) and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR), 2015 12th Latin American (2015) Oct 29; IEEE.

28. Latombe, Jean-Claude. "Motion planning: A journey of robots, molecules, digital actors, and other artifacts." The International Journal of Robotics Research 18, 11 (1999).
29. Marble, James D., and Kostas E. Bekris. "Asymptotically near-optimal planning with probabilistic roadmap spanners." IEEE Transactions on Robotics 29, 2 (2013).
30. Chung JJ, Lawrance N, Gan SK, Xu Z, Fitch R, Sukkarieh S. Variable Density PRM Waypoint Generation and Connection Radii for Energy-Efficient Flight through Wind Fields. InProceedings of IEEE International Conference on Robotics and Automation (2015).
31. https://en.wikipedia.org/wiki/Bio-inspired_computing, (2017), October 22, 2 PM.
32. https://en.wikipedia.org/wiki/Genetic_algorithm, (2017) August 07, 12:41 AM.
33. Li, Deshi, Xiaoliang Wang, and Tao Sun. "Energy-optimal coverage path planning on topographic map for environment survey with unmanned aerial vehicles." Electronics Letters 52, 9 (2016).
34. Dogru S, Marques L. Energy efficient coverage path planning for autonomous mobile robots on 3D terrain. InAutonomous Robot Systems and Competitions (ICARSC), 2015 IEEE International Conference on (2015) Apr 8 IEEE.
35. Tokuta, Alade. "Extending the VGRAPH algorithm for robot path planning." (1998).

Wireless Wearable for Sign Language Translator Device using Intel UP Squared (UP^2) Board

Tan Ching Phing[1], Radzi Ambar[1], Aslina Baharum[2], Hazwaj Mhd Poad[1] and Mohd Helmy Abd Wahab[1]

[1] Department of Computer Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[2] Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia
aradzi@uthm.edu.my

Abstract. Sign language translator devices translate hand gestures into text or voice that allow interactive communication between deaf and hearing people without the reliance on human interpreters. The main focus of this work is the development of a wireless wearable device for a sign language translator using an Intel UP Squared (UP^2) board. The developed device is consists of a wearable glove-based wearable and a display device using an Intel UP^2 board. When hand gestures have been created by a user, the accelerometer and flex sensors in the wearable are able to measure the gestures and conveyed the data to an Arduino Nano microcontroller. The microcontroller translates the gestures into text, and then transmits it wirelessly to the UP^2 board, subsequently displays the text on an LCD. In this article, the developed hardware, circuit diagrams as well as the preliminary experimental results are presented, showing the performance of the device, while demonstrating how the Intel UP^2 board can be connected to a low-cost Arduino microcontroller wirelessly via Bluetooth communication.

Keywords: Wearable Device, Sign Language Translator, Accelerometer, Flex Sensor, Intel UP Squared Board

1 Introduction

Deaf or hearing impairment person is someone who is unable to hear sounds either totally or partially. Deaf person usually use lip-reading, pen and paper, an interpreter or sign language to convey their thoughts and express their feelings. World Federation of Deaf stated that, there are 70 million deaf people in the world who use sign language as their first language or mother tongue [1]. Deaf person in Malaysia is using Malaysian Sign Language (MSL) that consists of dialect from different state. According to the Malaysian Federation of the Deaf, there are currently 32,157 persons with impaired hearing in Malaysia, but there are less than 100 certified sign language interpreters in the country [2-3]. The limited amount of interpreters are not sufficient to provide service for the deaf communities in the country [4]. Besides, both deaf community and public are quite dependent on sign language interpreter [5]. To avoid over-relying on limited amount of interpreter, sign language translator device has

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_52

been invented to eliminate communication barrier between deaf communities and public.

Nowadays, devices that can translate sign languages into voice and text have been researched and developed extensively. Basically, there are two methods that are usually used in studies related to sign language translator which are vision-based system and wearable devices. Vision-based systems utilize image processing method through feature extraction techniques to identify hand and finger movements [6-8]. On the other hand, wearable devices for sign language recognition usually utilize sensors attached on the user or glove-based approach [9-13]. Vision-based systems may not involve wearing sensory devices that can be uncomfortable, but, the system is complex, expensive and usually not portable. On the contrary, wearable devices can easily be a portable system and low-cost. But, wearables usually consist of data cables that limits hand movements. Therefore, a wireless wearable can easily permits free hand movement that can be portable and reduce costs. There are several works on the development of glove-based sign language translation device. Bui et al. developed a glove-based sensing device consists of six accelerometers attached on five fingers and back of the palm [9]. The collected data is processed by a low-cost Parallax Basic Stamp microcontroller, and then fed to a PC via a serial cable. The device recognizes about 23 Vietnamese-based letters. Jingqiu et al. invented a data glove consists of five (5) flex sensors and and ARM microprocessor that is connected to a display and audio module via cables [10]. Rishikanth et al. developed a low-cost sensor glove for gesture recognition consists of flex sensors and contact sensors [11]. The data from sensors are conveyed to an Arduino microcontroller via cables, and the text of the translated gestures are displayed on an LCD. Gałka et al. proposed an inertial motion sensing glove with gesture recognition method based on Hidden Markov Model [12]. The sensing glove consists of six (6) accelerometers to sense hand movements. The data are processed by a microcontroller and transmitted to a PC via a universal serial bus (USB) cable. The authors also have been developing a glove-based sign language translator device [13]. The described glove-based devices, however, are connected to display unit via cable system that limits free hand movements. Furthermore, the gloves utilized low-cost microcontrollers that are frequently being used in researches.

In this work, a wireless wearable device have been developed for a sign language translator system that translates sign language into text. The glove-based wearable consists of five (5) flex sensors, an accelerometer and an Arduino Nano, is wirelessly connected via Bluetooth communication to an Intel UP Squared (UP^2) board that displays the translated sign language on an LCD. This paper also describes the method on how to interface the high-end UP^2 board with a low-cost microcontroller.

2 Methodology

Fig. 1 shows an overview of the proposed sign language translator device which is consists of a glove-based wearable that is connected wirelessly via Bluetooth communication to an Intel UP^2 board. As shown in the figure, the wearable is consists of five flex sensors, an accelerometer sensor, an Arduino Nano, and an AT-09 BLE 4.0 module. The UP^2 board is connected to a BLE 4.0 module and a Grove-LCD RGB Backlight.

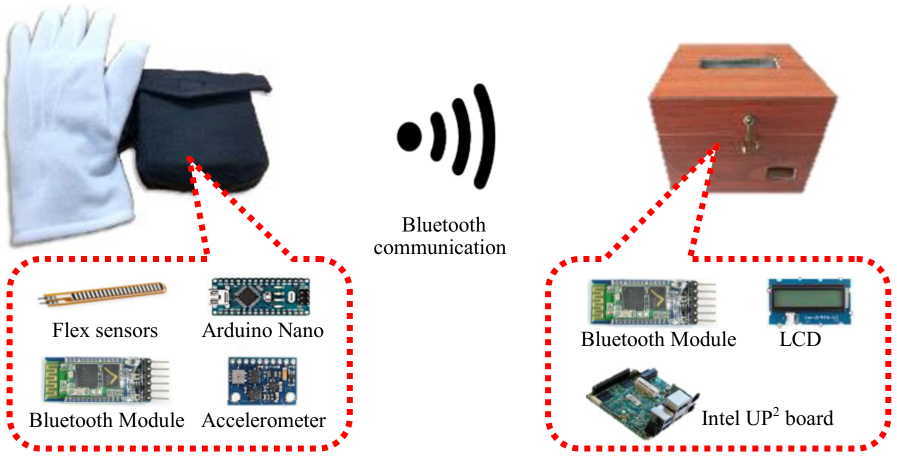


Fig. 1. Overview of the sign language translator device using Intel UP^2 board

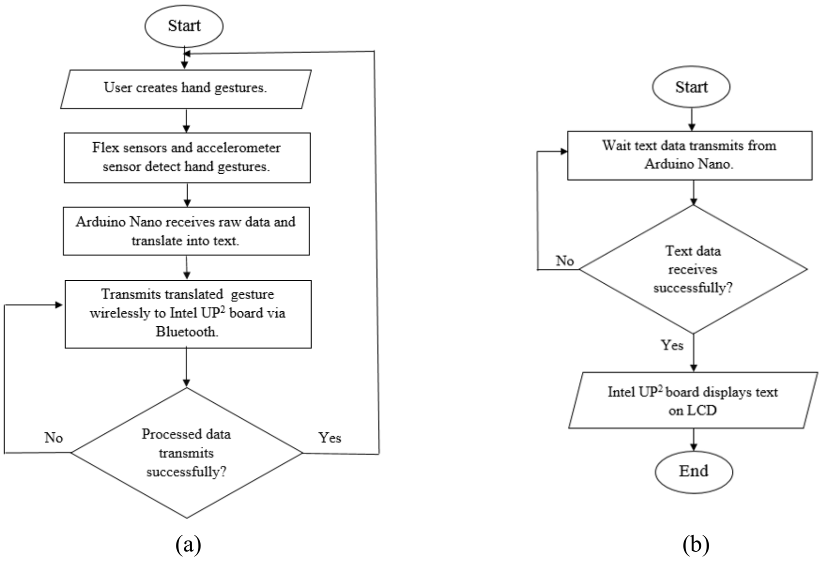

Start
User creates hand gestures.
Flex sensors and accelerometer sensor detect hand gestures.
Arduino Nano receives raw data and translate into text.
Transmits translated gesture wirelessly to Intel UP² board via Bluetooth.
Processed data transmits successfully?
No
Yes
(a)
Start
Wait text data transmits from Arduino Nano.
Text data receives successfully?
No
Yes
Intel UP² board displays text on LCD
End
(b)

Fig. 2. Flow chart for (a) wearable device and, (b) display device

Fig. 2(a) shows the flow chart of the wearable device. The wearable is able to read movements of wrist and every finger using the sensors. When the user creates a sign language gesture, Arduino Nano receives the raw data from those sensors. Arduino Nano processes the raw data from sensors, then translate it into text, and transmits the processed data wirelessly from Arduino Nano to UP^2 board via a Bluetooth connection. If the translated gesture transmitted successfully to the UP^2 board, the system will move back to “start” state and wait for another hand gestures input. At the same time, the UP^2 board displays the translated data in text on a Grove-LCD RGB Backlight. However, if the text data transmission failed, the system will continuously monitor the data transmission status. Fig. 2(b) shows the flow chart of the display device consists of the UP^2 board. The system will wait for the text data to be transmitted from Arduino Nano. If the system successfully receives data, UP^2 board displays the text on the LCD. Else, it will continues to wait data to be transmitted from Arduino Nano.

2.1 Circuit Diagram for Wearable Device

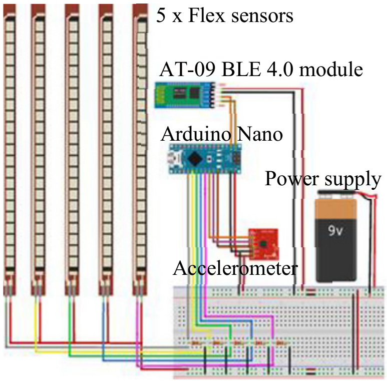


Fig. 3. Circuit diagram for the wearable device

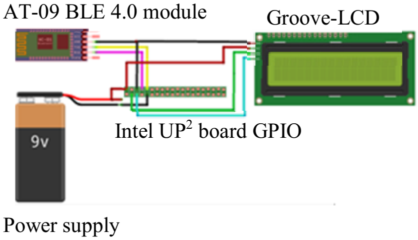


Fig. 4. Display device circuit diagram

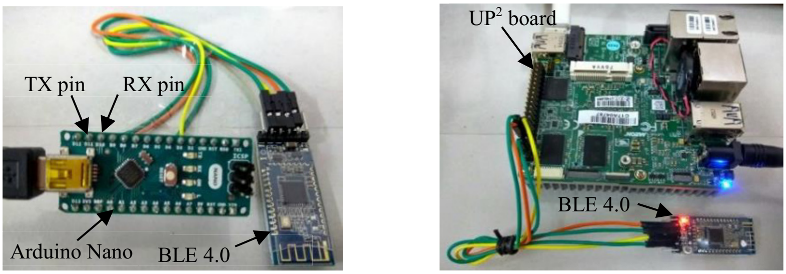
Fig. 3 shows the circuit diagram for the wearable glove-based device. The wearable is consists of an Arduino Nano, five (5) flex sensors, an accelerometer, an AT-09 Bluetooth Light Energy (BLE) 4.0 module, resistor 220Ω, jumper wires, power bank and a breadboard. Fig. 4 shows the circuit diagram for the display device that consists of the Intel UP^2 board. This system consists of an Intel UP Squared board, an AT-09 BLE 4.0 module, a Grove-LCD RGB Backlight, power supply and jumper wires.

2.2 Intel UP^2 board

Intel UP^2 board that is used in this project is one of the board in Intel’s UP board family. Previously, UP board is founded from Kickstarter campaign and it is has delivered to more than 12,000 makers for projects in Internet of things (IoT), industrial automation, home automation and digital signage [14]. Later, due to an overwhelming responds and supports from makers, UP^2 board is created with credit card size that has an ultra-compact single board computer with high performance and low power consumption. This board is based on the latest Intel® Apollo Lake Celeron™ and Pentium™ Processors with only 4W of Scenario Design Power and a powerful and flexible Intel® FPGA Altera MAX 10 on-board. The board is compatible with operating system such as Linux, Android and Windows 10. It comes with 2GB/4GB/8GB LPDDR4, 32GB/64GB/128GB eMMC, USB 3.0 ports, 2GB Ethernet and HDMI ports. It has the same form factor with Raspberry Pi that contains 40-pin GP-bus which provides the freedom for makers to build up their module. Additionally, there is a 60-pin EXHAT for embedded applications. This allows for the exploration of more possibilities. This project demonstrates the method to connect UP^2 board to a low-cost Arduino Nano wirelessly via Bluetooth communication.

2.3 Wireless Connection Setup Between Wearable and UP^2 board

As shown in Figs. 3 and 4, AT-09 BLE 4.0 modules are used to wirelessly connect the wearable to the UP^2 board. In order to set-up Bluetooth communication, both BLE 4.0 modules need to be configured either as a slave or master. First, the BLE 4.0 module in the wearable is set as slave. Fig. 5(a) shows the hardware connection between Arduino Nano and slave BLE 4.0 module in the wearable. As shown in the circuit diagram in Fig. 3, the TX and RX pins for BLE 4.0 module are connected to digital pin 10 and pin 11 of Arduino Nano. While the VCC and GND pins for BLE 4.0 module are connected to VCC and GND pin of Arduino Nano.

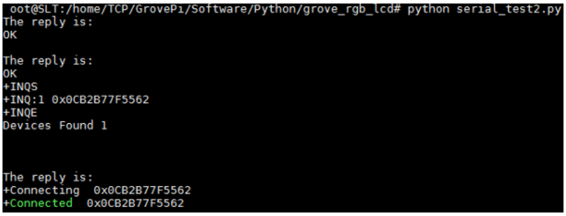


(a) (b)

Fig. 5. Hardware connection between BLE 4.0 module with (a) Arduino Nano, & (b) UP^2 board

Next, the BLE 4.0 module that is connected to UP^2 board is set as master. Fig. 5(b) shows the hardware connection between UP^2 board and master BLE 4.0 module. The TXD and RXD pins for master BLE 4.0 module are connected to UART_TXD and UART_RXD pins of UP^2 board. While the VCC and GND pins for master BLE 4.0 module are connected to 3.3V and GROUND pins of UP^2 board. In this work, the UP^2 board runs on Ubilinux operating system. The UP^2 board is a single board computer that requires peripherals such as a monitor display, keyboard and mouse to be connected to the board. As an alternative, for the ease of setup, a laptop installed with a software named MobaXterm was used as remote display to control the UP^2 board. There are few settings on MobaXterm before the board can be used including executing a Python source code that consists of three (3) Attention (AT) commands that is used to establish Bluetooth communication. The code is written in Nano text editor on the UP^2 Board via MobaXterm. The three AT commands that are used for Bluetooth communication are “AT”, “AT+INQ” and “AT+CONN1”.

Once the settings in Fig. 5 are executed, Bluetooth communication can be established. However, both the Arduino Nano and UP^2 board needed to be powered ON before establishing Bluetooth communication. Fig. 6 shows the response received at UP^2 board terminal when both BLE 4.0 modules are connected successfully. Firstly, “AT” command is used to check status of the master BLE 4. 0 module. It will reply “Ok” which means it is in AT command mode, and is not paired with other BLE 4.0 module. Next, “AT+INQ” command is used to scan nearby slave BLE 4.0 module. As shown in Fig. 6, the master BLE 4.0 module has found other BLE 4.0 module with the address of 0C:B2:B7:7F:55:62. This is the address for slave BLE 4.0 module that is used in the wearable. Then, “AT+CONN1” command is used to connect with the scanned slave module. Lastly, the Bluetooth communication is now established. In the next section, the experimental setup and steps to send data from the wearable to UP^2 board will be described.



```
 oot@SLT:/home/TCP/GrovePi/Software/Python/grove_rgb_lcd# python serial_test2.py
The reply is:
OK

The reply is:
OK
+INQS
+INQ:1 0x0CB2B77F5562
+INQE
Devices Found 1



The reply is:
+Connecting  0x0CB2B77F5562
+Connected  0x0CB2B77F5562
```

Fig. 6. Response received at terminal when both BLE 4.0 modules are connected

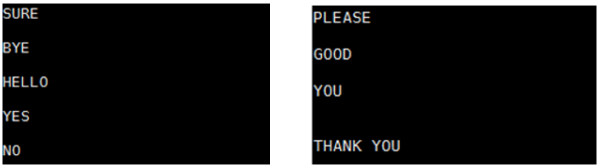
3 Experimental Setup and Results

3.1 Sending Text from Wearable to UP^2 board

After establishing Bluetooth communication using BLE 4.0 modules as in the previous section, data (translated gesture in the form of text) can be sent from the wearable to UP^2 board. A test source code for transmitting data has been written using Arduino IDE for Arduino Nano. Furthermore, a Python source code for receiving the data has been written using Nano text editor on the UP^2 board, and it is ran on the board’s terminal. Fig. 7 shows the results of several texts displayed on the UP^2 board terminal that have been transmitted from Arduino Nano via Bluetooth communication. From the figure, it shows that Bluetooth communication using BLE 4.0 modules between the wearable and UP^2 board are well established and data transmission was successful.

3.2 Display text on Grove-LCD RGB Backlight

A Grove-LCD RGB Backlight is used to display the translated sign language gesture received from the wearable. The LCD is connected to the UP^2 board as shown in Fig. 8. Initially, the LCD was tested to display “Hello world” as in the figure using Python code. The figure also shows that the UP^2 board was successfully interfaced with the Grove-LCD RGB Backlight.



```
SURE

BYE

HELLO

YES

NO
```

```
PLEASE

GOOD

YOU

THANK YOU
```

Fig. 7. Up^2 board displaying text data received from Arduino Nano at terminal

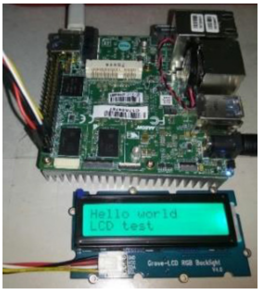


Fig. 8. "Hello world" was successfully displayed on the Grove-LCD RGB Backlight

3.3 Experiment on the Developed Sign Language Translator Device

After establishing Bluetooth communication between the wearable and UP^2 board, and making sure that the LCD has been successfully connected to the UP^2 board, experiment to show the integration of both wearable and display device has been carried out.

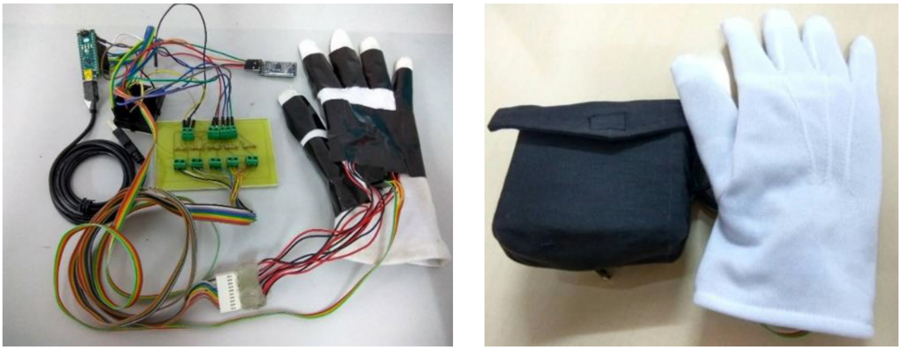


Fig. 9. The glove-based wearable. Hardware components (left), and end product (right)

Fig. 9 shows the developed glove-based wearable. The figure on the left shows the hardware components of the wearable consists of Arduino Nano, flex sensors, accelerometer, slave BLE 4.0 module, voltage divider circuit, bread board and jumper wires. The figure on the right shows the wearable end product with a small pouch bag containing the hardware which makes it portable and easy to carry.

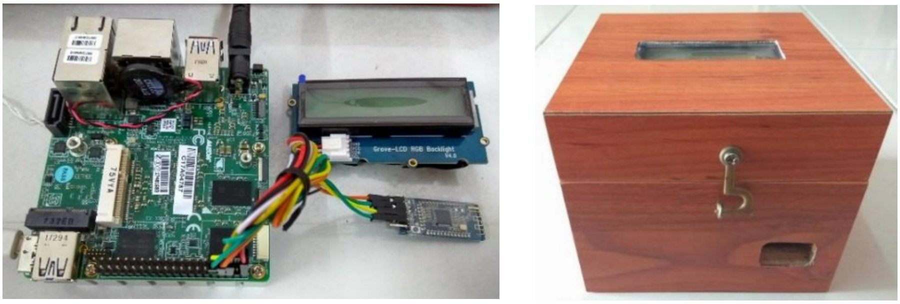


Fig. 10. The developed display device consists of UP^2 board and Groove-LCD (left), and a wooden case to store the hardware (right)

Fig. 10 shows the display device for the sign language translator device. The figure on the left shows the hardware of the device consists of an UP^2 board, a BLE 4.0 module, a Grove-LCD and jumper wires. The figure on the right shows a wooden case that is used to store the hardware of the device. Few holes were made to enable the usage of components such as Micro-USB cable, power adapter cable and Grove-LCD RGB Backlight.

Experimental steps. In order to show the usefulness of the developed device, settings described in subsection 2.3 were executed. Once Bluetooth communication is established, a normal, healthy user was asked to wear the wearable on his right hand. Then, the user was asked to execute nine (9) different sign language gestures which are "THANK YOU", "HELLO", "BYE", "YES", "SURE", "NO", "YOU", "GOOD" and "PLEASE". The image of the results displayed on the Grove-LCD is captured.

Experimental results and discussions. Fig. 11 shows the results of the experiment. The figure shows the sign language gestures and the corresponding translated texts displayed on the Grove-LCD. As shown in the figure, the gestures are "THANK YOU", "BYE", "YES", "NO", "YOU", "HELLO", "SURE", "GOOD" and "PLEASE". The results show that the developed sign language translator device has been successfully translated nine (9) basic sign language gestures using the wearable.

In order to determine the accuracy of this device, each gesture was tested for 50 times. Fig. 12 shows the result of the tests. Table 1 shows the number of error accumulated for each gesture. The device was able to translate sign language gestures for "THANK YOU" and "YOU" with 98% accuracy which are the highest. Only one error occurred for both gestures because of the position and direction of hand movement for them are unique compared with other gestures. However, the translation accuracy of sign language gestures for "YES" and "PLEASE" are 84%, which are the lowest accuracy result. The translation accuracy for sign language gestures like "BYE", "HELLO" and "PLEASE" are almost similar due to the fact that these three gestures have almost similar position and direction of hand movement. The accuracy for sign language gesture "YES" was low because it is difficult to bend all of the fingers simultaneously. Furthermore, the accuracy for sign language gesture "PLEASE" was low because this gesture was quite similar with sign language gesture for "HELLO".

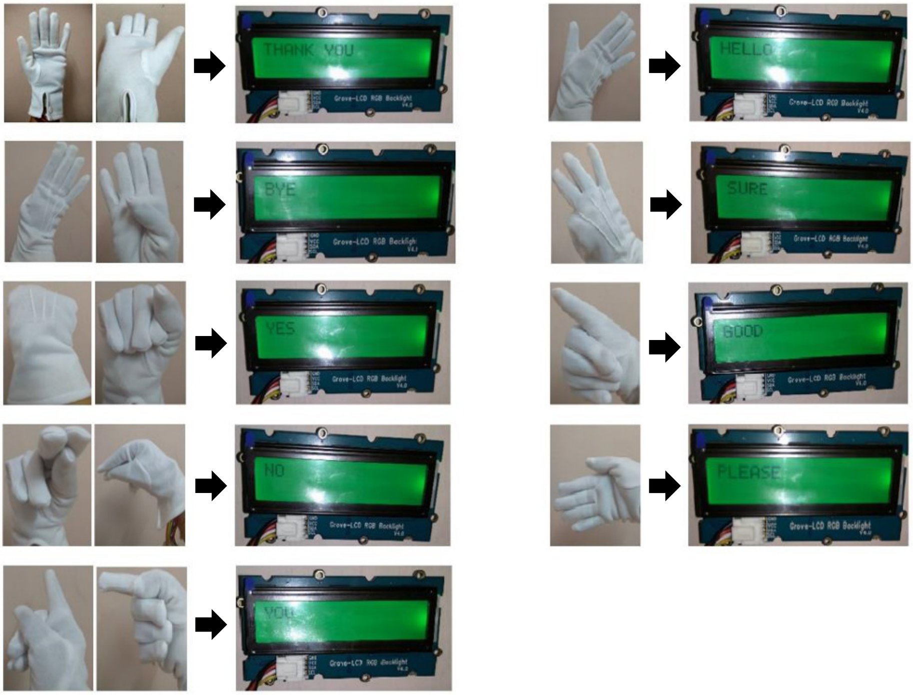
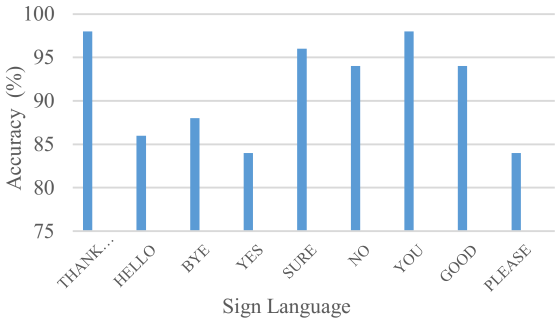


Fig. 11. Nine types of sign language gestures and the corresponding translated texts that are displayed on the Grove-LCD



Sign Language	Number of Error
THANK YOU	1
HELLO	7
BYE	6
YES	8
SURE	2
NO	3
YOU	1
GOOD	3
PLEASE	8

Fig. 12. Experiment results for accuracy test

4 CONCLUSION

In this paper, the design of a wireless wearable for a sign language translator device using Intel UP^2 board is described. The developed wearable has been successfully connected wirelessly to a display device containing an UP^2 board via Bluetooth communication. The paper also described the method to connect a low-cost Arduino Nano to a high-end Intel UP^2 board via Bluetooth communication. The experiment results presented in Section 3 shows that the developed sign language translator device is able to translate sign language gestures into text, and displays it on an LCD successfully. The proposed glove-based wireless wearable design permits free hand movements compared to a wired glove-based device. However, the display device in this work is still bulky and troublesome to be carried by a deaf person. As for future work, taking into account the ease-of-use, the wearable will be connected wirelessly to other devices such as a smartphone. A smartphone app will be developed to display text and produce sound of sign language gestures. This work provides a good learning curve in developing a wireless system using Bluetooth communication. Therefore, this project is only a small step for a larger ambitious project.

References

1. World Federation of the deaf (2016). Sign Language. https://www.malaysiakini.com/news/376165
2. Acute shortage of sign language interpreters in M'sia. https://www.malaysiakini.com/news/376165
3. MFD: Massive shortage of sign language interpreters. https://www.thestar.com.my/news/nation/2013/09/20/more-interpreters-needed/
4. Mohid, S. Z., Zin, N. A. M. (2011) Accessible courseware for kids with hearing impaired (MudahKiu): A preliminary analysis. Proc. 2011 Int. Conf. Pattern Anal. Intell. Robot. ICPAIR 2011, vol. 2, no. June, pp. 197–202
5. Simons, G.F., and Fennig, C.D. Ethnologue: Languages of the World, Twentieth edition. Dallas, Texas: SIL International. http://www.ethnologue.com.
6. Igari, S., Fukumura, N. (2016) Recognition of Japanese Sign Language Words Represented by Both Arms using Multi-Stream HMMs. In Proceedings of IMCIC-ICSIT, pp.157–162
7. Bauer, B., Kraiss, K. F. (2002) Video-Based Sign Recognition Using Self-Organizing Subunit. In Proceedings of Int. Conf. on Pattern Recognition, vol. 2, pp. 434–437
8. Dreuw, P. et al. (2007) Speech Recognition Techniques for a Sign Language Recognition System. InterSpeech-2007, pp. 2513–2516
9. Bui, T. D., Nguyen, L. T. (2007) Recognizing Postures in Vietnamese Sign Language With MEMS Accelerometers. IEEE Sensors Journal, vol. 7, no. 5, pp. 707–712
10. Jingqiu, W., Ting, Z. (2014) An ARM-Based Embedded Gesture Recognition System Using a Data Glove. The 26th Chinese Control and Decision Conference, pp. 1580-1584
11. Rishikanth, C. et al. (2014) Low-cost intelligent gesture recognition engine for audio-vocally impaired individuals. 2014 IEEE Global Humanitarian Technology Conference (GHTC), pp. 628-634
12. Gałka, J. et al. (2016) Inertial Motion Sensing Glove for Sign Language Gesture Acquisition and Recognition.IEEE Sensors Journal, Vol. 16, No. 16, pp. 6310-6316
13. Ambar, R. et al. (2018) Preliminary Design of a Dual-Sensor Based Sign Language Translator Device. In: Ghazali R., Deris M., Nawi N., Abawajy J. (eds) Recent Advances on Soft Computing and Data Mining. SCDM 2018. Advances in Intelligent Systems and Computing, vol. 700, pp. 353-362

14. UP^2 (UP Squared) Unveiled On Kickstarter. http://www.up-board.org/up/press-release/up%C2%B2-up-squared-unveiled-on-kickstarter/

APTGuard : Advanced Persistent Threat (APT) Detections and Predictions using Android Smartphone

Bernard Lee Jin Chuan, Manmeet Mahinderjit Singh, Azizul Rahman Mohd Shariff

School of Computer Sciences,Universiti Sains Malaysia
bernard.ucom13@student.usm.my; {manmeet; azizulrahman }@usm.my

Abstract. Advanced Persistent Threat (APT) is an attack aim to damage the system's data from the aspect of confidentiality and integrity. APT attack has several variants of attacks such social engineering techniques via spear phishing, watering hole and whaling. APTGuard exhibits the ability to predict spear phishing URLs accurately using ensemble learning that combines decision tree and neural network. The URL is obtained from the SMS content received on the smart phones and sent to the server for filtering, classifying, logging and finally informing the administrator of the classification outcome. APTGuard can predict and detect APT from spear phishing but it does not have the ability of automated intervention on the user receiving the spear phishing URL. As a result, APTGuard is capable to extract the features of the URL and then classify it accordingly using ensemble learner which combines decision tree and neural network accurately.

Keywords: Advanced Persistent Threat (APT); Spear-phishing; Ensemble Learning; Data Mining; Neural Network

1 Introduction

Advanced Persistent Threat (APT), has been hitting everyone retaining valuable data, either government organizations or non-government organizations such as corporations, business firms and many more for a long period with cutting-edge techniques that are so subtle and remained undetected for a long time [1]. APT aims to exile the victims' data [2], be it valuable or not at that moment, which will be valuable when combined with other stolen data. When APT has been detected, the damage to the data, especially from the aspect of confidentiality has already been done [3]. Outmoded cyber security protocols such as firewall, intrusion detection and prevention system (IDPS), or antivirus are not able to defend against APT as it employs social engineering, which is to exploit the ignorance of unwary human beings into giving access for the attackers [4]. Now, cyber security firms are now focused onto predictive analytics as APT attacks changes every time and so subtle until conventional cyber threat defense are rendered useless when facing APT [5].

A well-staged spear phishing attack can lead to watering hole and malware infection. Thus, it is wise for us to target to curb the spear phishing. However, current cyber defenses are unable to defend against spear phishing as the tactics used to spear

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_53

phish is social engineering. Social engineering is that the users are tricked to giving out certain information such as credentials of a system. Spear phishing will cause the people especially for the corporate world some major financial and reputation damage [6]. The attack itself has cause the company suffering some losses and legal claims from some employees. APT always comes in the form of spear phishing which is a type of URL based attack. Some measures have been taken to stop spear phishing. The first defense against spear phishing is users are well trained to recognize these attempts [7-9]. However, the issue still occurs as the saying “To err is human” goes – Human tends to make mistakes. Since human do make mistakes, then web filtering, signature-based security tools such as firewalls, IDPS is used in attempt to stop spear phishing. [10] However, the issue still occurs due to the ever-changing attack vector that newly-emerge every day. All new attacking vectors will be unrecognized by the signature-based security tools. Therefore, a system that predicts that possibility of spear phishing will be able to prevent such situation occurs. Thus in this paper, we proposed APTGuard which aims to classify the URL by using decision tree and neural network based on the characteristic of the URL itself. With ensemble learning, URL can be classified without user intervention and indirectly, spear phishing can be detected and predicted.

2 Background: Advanced Persistent Threats (APT) Overview

Mobile devices have been a target for APT hackers as network administrators have little or no control on it [11]. Per the rule of the bucket, information security is only as strong as its weakest link and in this case, mobile devices are the weakest link where APT attackers made their moves. In this section, APT attack will be discussed in depth.

APT targets on the victim’s digital property that bring competitive or strategic benefits while traditional threats looking for personal data that provide financial rewards [13]. APT always have a group of highly planned, well-resourced actors [13]. The “P” in the term APT stands for persistent. It means that the attack will keep going on until it is successfully done. Apart from traditional threat, traditional threat only attack for some time and will move on to the next target which less secure if they failed to attack the current ones. Finally, APT is stealthy and evasive [12]. The attack stays in a low profile as long as they can accomplish their objectives. APT attack model usually consists of 6 phases: Reconnaissance and Weaponization, Delivery, Initial Intrusion, Command and Control, Lateral Movement and lastly, Data Exfiltration [14].

3 Proposed Solution: APTGuard

APTGuard is using the combination of smartphone sensors to collect the ambient environment of the user which can predict the condition of the user and his/her smartphones and finally, analyse the possibility of an APT attack. As mentioned in [11, 12, 14], IDPS has a high rate of false positive or false negative. Hence, the proposed system will be embedded with predictive intelligence analysing fusion of sensor data from an Android smartphone based on user behavior.

Table 1. Test Cases for the Fusion of Smartphones Sensors

Scenario	Sensors	Process
The user receives an SMS.	SMS	Mobile application obtains and sends the SMS content to the server for processing. The ensemble learner will then return the result and save it to the database.

Based on Table 1, if there is a URL contained in the SMS, the URL will be classified using ensemble learner. The result will then be logged into the database. If the classification outcome claims that the URL is malicious, an alert will be sent to the administrators. The system should be capable of:

- Predicting the possibility of URL spear phishing by employing a set of preprocessing features to determine the characteristic of the URL.
- Detecting and predicting spear phishing URL by using ensemble learner.
- Providing logging mechanism based on notification and alerts from the engine.

The system will be able to check on the features of a URL which are, Similarity against Browser History, Information in Google Safe Browsing, Web crawling for the Content Source, Google Domain Search, Domain Age, Alexa Ranking and Google Search. These features are chosen as the pre-processed data for the ensemble learner to classify the URLs. The features are chosen on the basis as of below:

Table 2. Chosen Features of a URL and its Justification

Feature	Justification
Similarity against Browser History	Comparison against the browser history to check the existence of similar URL in the browser history then a comparison of URL similarity is made. This is done because spear phishing tends to have a similar URL compared to the genuine site to deceit the victims.
Information in Google Safe Browsing	The URL is compared to the database list in Google Safe Browsing to make a direct checking on the URL.
Web crawling for the Content Source	Web Crawling is done to check for the content source of the URL. Spear phishing sites tend to use content from genuine sites so that the site can be more “believable”.
Google Domain Search	The URL domain will be checked for the existence in Google.
Domain Age	Domain age for spear phishing URLs will be shorter as the site was recently built to deceit its victim.
Alexa Ranking	The ranking of the site will be checked as the spear phishing site tend to imitate from popular sites and popular sites will have a higher ranking and the imitate site will have a much lower ranking.
Google Search	Google will hide all the spear phishing site from its search engine.

The ensemble learner will compare all the features stated on Table 2. The ensemble learner is the combination of a decision tree and convolutional neural network. The URL and its features will then put into the ensemble learner to classify the URL to determine whether the URL is good or malicious.

4 APTGuard System Architecture

This project is to be used by the users from the corporates which information is their assets. There are 2 actors which are users, which are required to install the client application to access to the corporate network and the network administrators, which will be the group that receives notifications from the system regarding the possibility of an attack. Fig. 1 shows the low-level system architecture diagram for APTGuard.

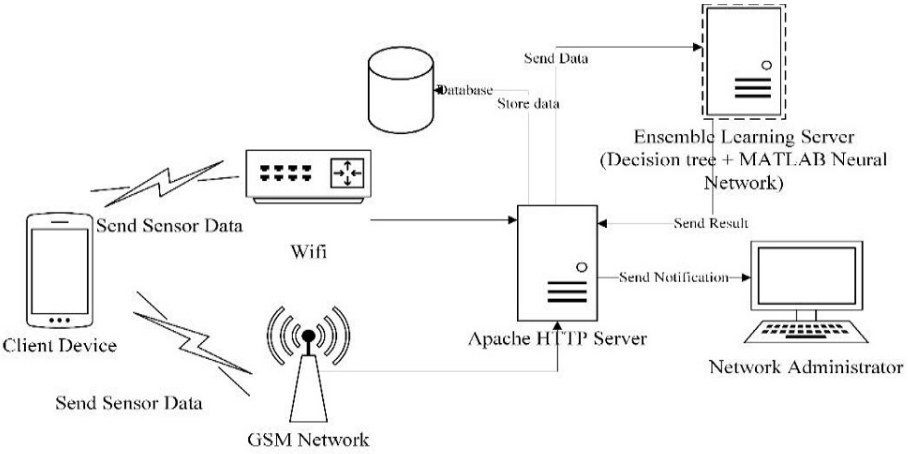


Fig. 1. APTGuard System Architecture Diagram

Based on the diagram, the client app, which will be installed on client's devices will be employing Wi-Fi (IEEE 802.11x) or Cellular Network to send smart devices sensor data to the server. At the server side, there will be a database and an ensemble learning server which will be used to store data and performing ensemble learner classification respectively. The server will also send notification to the user if there is a possibility of an URL-based APT attack. The server will also enable the administrator to check on the log file by searching through the database via server.

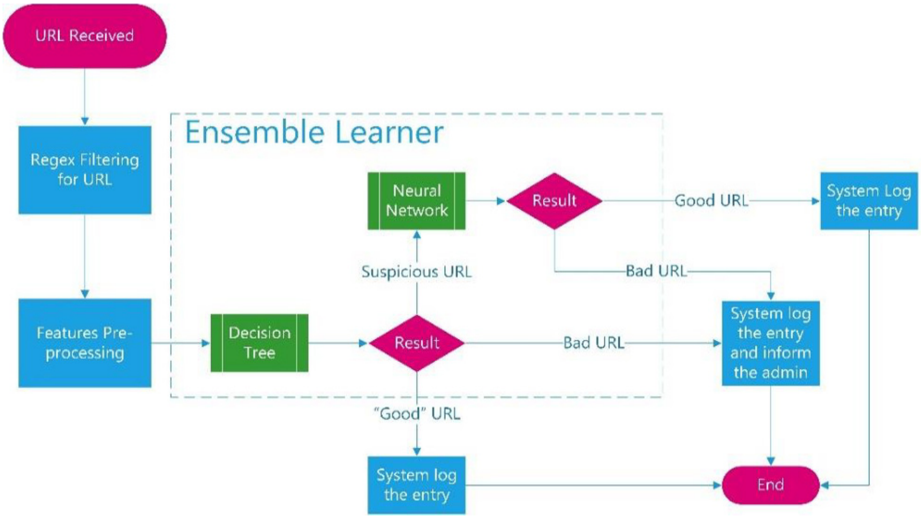


Fig. 2. Flow Chart of APTGuard

Based on Fig. 2 shown, the user will install the application into their phone and must give full permission to the applications to activate it. Once the client application is activated, it will start sending data to the server when there is an SMS received from the user smart phones. The servers will send the data collected to the server for ensemble learner classification and then store it to the database and if the ensemble learner outcome is positive on URL-based APT attacks, the network administrators will receive an alert on their computer or console regarding the incoming attack.

5 APTGuard Design And Implementation

The users of APTGuard are divided into 2 groups which they are:

- System Administrator: Administrator is the users who have the most privileges and authorisation. The admin can create a new account for the new onboarding administrators and employees so they have the access to APTGuard. The admin will be able to login to the web application to perform various tasks such as creating new users, checking on the neural network metrics such as sensitivity, confusion matrix.
- Employees: An employee is the "ordinary user" in APTGuard. They can only provide data for the system to detect possibility of spear phishing.

5.1 Web Application

Web application is to be used by the administrators. The web application is implemented using HTML, CSS and PHP. It is hosted on an Apache HTTP Server, using MySQL Database. The purpose of creating a web application is to fulfil the purposes stated below:

System Dashboard. System Dashboard is where the administrators are informed of the status of APTGuard in a summary view. The aim of having a dashboard is to consolidate all vital information into 1 page so that the administrator will not left out some good-to-know heads up information. The system dashboard is viewable in the main page of the web application after the admin performed successful login. System dashboard will display some of the vital information such as number of spear phishing URL received from the user or the user receiving the most number of spear phishing URL.

Neural Network Metrics Information. Neural network is a machine learning that can be measured of its performance by using confusion matrix especially when dealing with the problem of statistical classification. A confusion matrix will be plotted in this part to be considered as a general information for the admin to monitor the performance of the neural network. Besides, a series of figures such as recall, precision, miss rate, fall-out and accuracy will be derived from the confusion matrix. The neural network information is obtained using Python.

5.2 Mobile Application

The mobile application is an Android application design specifically for the user. It serves the purpose of sending SMS received from the user devices to the server to the system backend for further processing. mobile application is implemented using JAVA and XML as the programming language and mark-up language respectively.When the user attempts login to the system, he/she is required to provide the correct and valid credentials. The credentials will be given by the system administrators. In the page, the user is required to provide credentials to authenticate himself/herself.

5.3 System Backend

The backend logic is the main part of the whole system. It processes any data received from the user devices and classify any URL found in the data received from the smart devices. The detailed responsibilities are stated below:

Receive data from the smart devices. When the user receives an SMS, the mobile application will send the content to the server. The server will be looking for any URL within the SMS content by performing regex searching. If URL is found within the content of the SMS, the URL will be classified using decision tree and neural network to determine whether the URL is malicious or not. Then the message and the results will be stored into the database. If there is no URL found in the SMS content, the message will then be stored into the database without and classification done.

Proposed Logic using Decision Tree Learner. Classification using decision tree means that there will be a series of conditional logics that the URL received will be going through. The conditions are listed in Table 4: Conditional Logics in APTGuard Decision Tree.

Table 4. Conditional Logics in APTGuard Decision Tree

Conditions	**Outcome**
URL Similarity against system URL history [13]	If similarity = 100% then not spear phishing.
Google Safe Browsing [15]	If result return bad then the URL is malicious
Web crawling for any external web content [15]	< 22% - Good URL 22%<x< 61% OR N/A - Suspicious > 61% - Bad
Google Domain Search [15]	If exist then it is good Else depending on domain age, Alexa ranking
Google Domain Search [15] & Domain Age [13]	If Domain Age > 6 months then suspicious Else URL is bad
Google Domain Search & Alexa Ranking [13]	If Alexa Ranking < 100000 then suspicious Else URL is bad
Google Search [13] & Domain Age [15]	If Exist && Age<6 months then suspicious
Google Search [13] & Alexa Ranking [15]	If Exist && Ranking <100000 then good If Exist && Ranking >100000 then suspicious If Ranking is not available then suspicious

The decision tree is implemented using PHP. PHP will be the main language that process the features extracted from the URL.

Classification using Neural Network. Convolutional neural network will be used to classify the URL if the decision tree classifies the URL as suspicious. The neural network will take in the 7 attributes, which are, similarity against the browser history, information in Google Safe Browsing, web crawling content for the content, Google Domain Search, Domain Age, Alexa Ranking and Google Search. Then, the outcome will be obtained from the neural network. The neural network is pre-trained before it can be used as a classifier. Since the neural network is part of the ensemble learner, the neural network will be invoked if the decision tree classifies the URL to “suspicious”. If the URL is suspicious, the ensemble learner invokes the neural network to classify the URL from suspicious to either “good” or “bad”. Then, the ensemble learner will store the result into the database.

6 System Testing And Evaluation

6.1 Testing Strategy

System testing has been carried out to ensure that the system remains functional, usable, reliable and efficient. The testing that has been done for APTGuard are, unit testing, integration testing and system testing. Unit testing means that the system is tested in the level of every elements of the system, which means that the test is conducted to verify the functionality of certain section of the code. In this case, every function/class in APTGuard will be tested to ensure that the functionality fulfils the objectives of the project. Unit testing is conducted throughout the development phase which is the test is carried out during the development stage to ensure that the code written works fine. System testing is done after all the modules are being integrated as the whole system. The test is only done after the development phase has been completed and the integration test has passed. System testing should be encompassing the design, system behavior and the expectations from the user.

6.2 Case Scenario Using APTGuard

Two case scenarios have been designed to fully test the system functionalities.

Case 1: - User receives an SMS with a URL within the content

Table 5: User receives an SMS with a URL within the content.

Step 1	User receives an SMS with a URL within the SMS content.
Step 2	Server receives the SMS from the mobile devices.
Step 3	Server performs regex to obtain URL within the SMS content.
Step 4	The obtained URL is being classified by the ensemble learning algorithm (decision tree + neural network)
Step 5	The result of the classification is being obtained and stored into the database.
Step 6	Administrator login to the web application to check on the result of the classification.
Step 7	Step 1 – 6 is repeated using good, bad and suspicious URL.

Case 2 – Experimental Result of APTGuard with Neural Network
In this section, the authors will provide the results of the experiment and discussion about it. There are total 1445 of sample URLs have been collected. The sources of the URLs are 499 from crazyurl, 412 from PhishTank, 500 from Alexa and the remaining from SMSs. From the samples, 856 URLs are bad websites and 589 URLs are good websites. Based on Table 6, the system achieves sensitivity rate of 91.00%, fall-out rate of 1.05%, precision of 98.35%, and accuracy of 95.71%.

Table 6. Result of neural network prediction model.

	Predicted Good	**Predicted Bad**	**Total**
Condition Good	536	53	589
Condition Bad	9	847	856
Total	545	900	1445

The system utilizes the neural network toolbox from the MATLAB tool. Its two-layer feedforward network which is a multi-layer feed forward network. Figure 3 shows about the architecture of the Neural Network System.

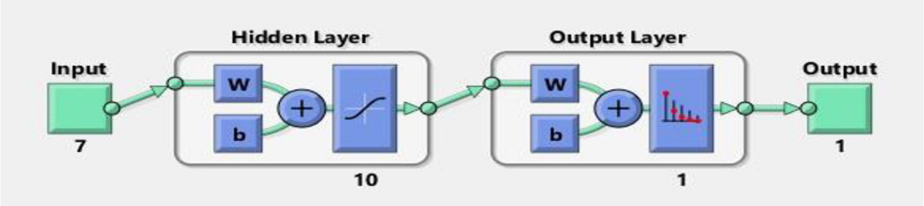


Fig. 3. MATLAB Neural Network System Architecture

All the data from the input layer will gone through the hidden layer with a sigmoid transfer function to produce output which then transform and calculate into 1 of the output in the output layer using softmax transfer function. There is only 1 type of output which represents the class representation of the URL.

6.3 APT Guard Security Requirements

ITU-T Recommendation X.800 has been used as the standard to define the security requirements of APTGuard. APTGuard will be evaluated according to the X.800 security standards. X.800 evaluates the system from the aspect of authentication, access control, data confidentiality, data integrity, non-repudiation. Authentication means that only the correct personnel with the correct credentials can access to the resources. This ensures that the resources can only be accessed if and only if the users and the administrators of the system provide the correct credentials before accessing any resources.

Access control restricts the access of the system and the resources within the system to certain parties, preventing unauthorized access into the system. In APTGuard, Bell-LaPadula Model is used to enforce access control – The administrators have full access of APTGuard while the access of a normal is only the subset of the administrator's access. Data confidentiality is the protection of information from unauthorized disclosure. The information should be hidden against unauthorized parties. Information such as personal information of the user and admin should be hidden from anyone except the person himself/herself and the administrators. The user and the admin password will be salted and hashed to prevent anyone from accessing it. Data Integrity is the assurance of the information obtained, stored are protected against fabrication. APTGuard will be employing password authentication ensuring only the authorized person can change the information that he has the right to change. Non-repudiation is the inability of a personnel to deny against any actions involved. APTGuard will record any information, action performed by the user or the admin.

7 Conclusion And Future Work

APT is a threat that cannot be predicted via conventional cyber security measures as it employs a series of social engineering techniques that tricks the user to provide the credentials to access into the system. This project covers the prediction of the possibility of an APT from mobile devices. It will obtain devices ambient sensor data and process it using predictive analytics and notify the admin before the threat happens, which mean the planned attack will be known even before it happens. These are the possible expansion of APTGuard such as i) APTGuard can extend its functionalities that perform automated intervention after a spear phishing attack is predicted; ii) APTGuard can extend its coverage that not only monitor the SMS content and APTGuard can employ multi-factor authentication that ensures only authorized personnel can access to the system.

Acknowledgement

This work is supported by USM-Maxis Grant No: 304/PKOMP/650804/M151

References

1. E., C. *Cyber Defense: Advanced Persistent Threat (APT) and Insider Threat*. 2012 [cited 2016 19 September 2016]; Available from: https://cyber-defense.sans.org/blog/2012/10/23/advanced-persistent-threat-apt-and-insider-threat.
2. Tankard, C., *Advanced persistent threats and how to monitor and deter them.* Network security, 2011. **2011**(8): p. 16-19.
3. FireEye. *Anatomy of an APT (Advanced Persistent Threat) Attack*. 2016 [cited 2016 20 September 2016]; Available from: https://www.fireeye.com/current-threats/anatomy-of-a-cyber-attack.html.
4. Sood, A.K. and R.J. Enbody, *Targeted cyberattacks: a superset of advanced persistent threats.* IEEE security & privacy, 2013. **11**(1): p. 54-61.
5. Lab, K. *Kaspersky Lab and SynerScope join forces to battle cyberthreats with big data analytics*. [cited 2016 22 September 2016]; Available from: http://www.kaspersky.co.in/about/news/business/2016/Kaspersky-Lab-and-SynerScope-join-forces-to-battle-cyberthreats-with-big-data-analytics. .
6. B., K. *KrebsOnSecurity*. [cited 2017 25 April 2017]; Available from: https://krebsonsecurity.com/2016/03/seagate-phish-exposes-all-employee-w-2s/. .
7. FireEye. *Best Defense Against Spear-Phishing*. [cited 2017 25 April 2017]; Available from: https://www.fireeye.com/current-threats/best-defense-against-spear-phishing-attacks.html. .
8. D., O. *SANS - Information Security Resources*. [cited 2016 19 September 2016]; Available from: https://www.sans.org/security-resources/idfaq/what-is-a-false-positive-and-why-are-false-positives-a-problem/2/8. .
9. D., O. *SANS - Information Security Resources*. [cited 2016 19 September 2016]; Available from: https://www.sans.org/security-resources/idfaq/what-is-a-false-negative/2/9.
10. Lin, C.-H., et al. *Efficient spear-phishing threat detection using hypervisor monitor*. in *Security Technology (ICCST), 2015 International Carnahan Conference on*. 2015. IEEE.
11. Research, D., *The Impact of Mobile Devices on Information Security: A Surveys of IT Professionals.* Dimensional Research, 2013.
12. Chen, P., L. Desmet, and C. Huygens. *A study on advanced persistent threats*. in *IFIP International Conference on Communications and Multimedia Security*. 2014. Springer.
13. Mohammad, R.M., F. Thabtah, and L. McCluskey, *Intelligent rule-based phishing websites classification.* IET Information Security, 2014. **8**(3): p. 153-160.

14. A., J. *Anatomy of an APT Attack: Step by Step Approach*. [cited 2017 25 April 2017]; Available from: http://resources.infosecinstitute.com/anatomy-of-an-apt-attack-step-by-step-approach/. .
15. Basnet, R.B., A.H. Sung, and Q. Liu. *Rule-based phishing attack detection*. in *Proceedings of the International Conference on Security and Management-SAM*. 2011.

Smart Home using Microelectromechanical Systems (MEMS) Sensor and Ambient Intelligences (SAHOMASI)

Manmeet Mahinderjit Singh[1], Yuto Lim[2], Asrulnizam Manaf[3]

[1]School of Computer Sciences,University Sains Malaysia (USM)
[2]Japan Advanced Institute of Sciences and Technology (JAIST)
[3]Collaborative Microelectronic Design Excellence Centre (CEDEC), University Sains Malaysia
manmeet@usm.my, ylim@jaist.ac.jp, eeasrulnizam@usm.my

Abstract. Smart home is a home that has a set of electrical appliances connected to a network can be remotely controlled, accessed and monitored. Smart home normally is used to fine-tune human's daily life. However, the main problem is when different devices such as sensors and electrical appliances produced by different manufacturers being used in a smart home. Each device produced by different manufacturer might come with their own set of communication protocol. Therefore, it is important to have a standardized communication protocol for interconnecting and remotely controlling those devices to ensure data can be transmitted to each other. This project aims to present the Smart Home using MEMS Sensor and Ambient Intelligence (SAHOMASI) where in this paper, a standardized communication protocol is proposed to overcome the above problem. The features of the system includes monitoring elderly's activities, detect fall and sending email to caregiver upon abnormal activity or fall detected.

Keywords: Smart Home, ECHONET Lite, Standardize Communication protocol, Internet of Thing(IoT), Ambient Intelligence, Ageing Society

1 Introduction

Smart home technologies have evolved over time and are implemented to help, ease and improve human life and reduce energy wastage (Energy Conservation). Mostly smart home technology will have ambient intelligence (Aml) integrated due to its non-intrusive way of interaction. This smart home often uses several ambient sensors at home and uses these sensors to collect data about human daily activities. Smart home will then try to learn and predict the pattern from the collected data for future uses. Smart home consists of 3 parts which are network, controlling devices and home automation [1]. The network is responsible for transmitting data from sensors to a controller and establishes connection between controlling devices and home automation, while controlling devices are responsible for managing how the system will work or response and lastly home automation is devices responsible for manipulating the physical environment [1].

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_54

There are few types of smart home developed around the world, for instance smart home developed in Egypt is used for energy saving due to Egyptian's habits that tend to forget to switch off any electrical appliances when are not in use [1]. Other than that, a smart home called Cloud-base smart home is developed to provide entertainment, security, environmental, domestic appliances, information and communication as well as health application [1].

Smart home in this world of technology is still considered to be very expensive and not everyone is affordable to implement this technology in their house. But this technology is very useful and helpful for elderly. Elderly with aging problems such as decline in memory, reasoning and communication skills and gradual loss of skills needed to carry out daily activities [2] required a lot of effort and resources to take care of them. According to a research, population of elderly aged greater than 65 in 2011 is about 17% and is estimated to reach 33% in year 2035 [2]. This situation will cause the ratio between elderly and caregiver to become imbalanced. With smart home its help to monitor elderly's daily activities and send alerts to respective caregivers if the system detects abnormal activities occurred, therefore its help in reduce the burden of the caregiver and help the elderly to live independently.

There is a problem that exists in current Smart Home design which is there are no standardize communication protocol used. Many current smart homes are using different kind of electrical appliances that produced by different manufacturer with different communication interfaces, therefore data management problem might arise due to different communication protocol being employed. Therefore, it is important to have a standardized communication protocol for interconnecting and remotely controlling those devices being used and to ensure data can be transmitted to each other. Besides that, it might cause difficulty in integrating existing application with another application in the future due to different communication protocol being used whereby one application must be changed to comply with another application communication protocol.

The overall objective of the system built is to set-up human movement activities within a smart living space using a standardized communication protocol with the use of these comparable cheap technologies such as Raspberry Pi, Arduino boards, a few types of sensor, and ECHONET Lite.

2 Background and Related Works

Smart home that proposed in papers [3-5] are implemented for monitoring the home, provide safety, convenience and comfort for the resident as well as for energy conservation purpose. The resident can control home appliances and monitoring their home when they are away no matter where and when. The implementation of smart homes that proposed in papers [3-5] are similar to each other, which they will use a series of sensors for sensing the activity and will have a node for collecting data from different devices including sensors and home appliances and is for data processing.

Layout the architecture of the smart home, which including a base station and a number of end node. Data will be transmitted wirelessly using radio technology between base station and end node [3]. Arduino board will be integrated with a microcontroller to act as a base station for data processing. Each end node will be responsible for data collection and send it to the base station. Killing et.al [3] have employed a number of technologies in their proposed system. This include RFID tag reader, Arduino cellular module, Infrared sensors, xBee radio, temperature sensors, smoke/CO detector, buzzer, voice recorder, security camera and a stepper motor.

Lin [4] has layout a few of consideration when we are implementing smart home which include the maintainability, future development and cost. Therefore Lin [4] has proposed smart home that is open source to overcome the issue. In Lin proposed smart home [4], it also uses an Arduino board to act as a controller for computing and carry out the overall system functionality, where Arduino board is an open source platform and is available to public and well-known by the public.

A research did by Piyare and Lee [5] has proposed low cost and flexible smart home controlling and monitoring system by using smart phone. They propose using an embedded micro-web server, with IP connectivity for accessing and controlling devices and appliances remotely using Android based smart phone.

2.1 Ambient Intelligence in Smart Home

Ambient intelligence is an emerging discipline that brings intelligence to our everyday environment and makes those environments sensitive to us [6]. Ambient intelligence (AmI) research builds upon advances in sensors, sensor networks, pervasive computing, and artificial intelligence. The overall idea behind AmI is that it will integrate the environment with full of technology (sensors and devices), so that sensing of activities, reasoning about the data collected, and act or take action based on certain criteria can be carried out by a system[7].

Ambient Intelligence Smart Home Automation (AMISHA) [8] system is one of the smart home developed to help the elderly live independently, to reduce the effort as a caregiver to take care of elderly, reduce the cost in term of hiring caregiver to be with elderly 24/7, and is developed in a relatively low cost compared to other existing smart home. AMISHA is designed with few sensors and these sensors will be connected to an Arduino with wireless transceiver. Sensor data will be transmitted to Arduino wirelessly and from the Arduino to laptop through an Ethernet Gateway. The laptop will then process the data and stored the process data in database and transformed into meaningful data and display in a web application for caregiver usage. AMISHA consists of three major modules: ambient sensors, prediction algorithm, and notification engine. The ambient sensors will be installed around the home and data will be collected and will be processed in prediction algorithm module to produce meaning data and information. iHome [9] is an extension to AMISHA, both provide similar functions, but fall detection algorithm has been implemented in iHome and is a cloud based application. In iHome, 3 modules have been implemented which include user module, server module as well as sensor and actuator module. User module will contain all the functions that are performed by users such as caregiver, elderly and admin. In server module, it is responsible for the overall system services and engines and is managed by the administrator. The sensor and actuator module contains all the sensors and actuators that used in iHome. Ambient sensors are developed using Arduino-based boards while the actuator is implemented using Raspberry Pi to simulate the real home appliances. Both smart home support for notifying elderly's caregivers upon abnormality detection through various mediums such as email, SMS, and Twitter.

2.2 Standardise Middleware : ECHONET Lite

ECHONET [10, 11] is stands for Energy Conservation and Homecare Network, that consist of a communication protocol that required no new wiring and can be installed in existing homes, multivendor-compatible home network equipment, and system models to facilitate the development of the application system. ECHONET Lite is developed due to the global environment issue such as global warming, and the need to reduce CO2 as well as the increasing cost for healthcare and growing need for nursing care. The ECHONET Lite design is based on ECHONET, that ECHONET Lite has removed ECHONET addresses, removed functions that are not implemented well in ECHONET as well as simplifying the message composition. ECHONET Lite is an open communication protocol that helps to support smart house and Home Energy Management System (HEMS). Since the smart home will integrate with different kinds of electrical appliances with different kinds of communication protocol, there is a need to absorb the differences of communication protocol, so that different kinds of electrical appliances can communicate with each other.

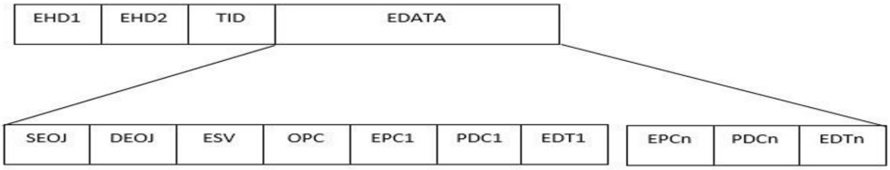


Fig. 1. Message Structure in ECHONET Lite

The above diagram shows the data packet used in ECHONET Lite communication protocol. EHD1 and EHD2 is stands for ECHONET Lite Header (EHD), TID stands for Transaction ID, and EDATA stands for ECHONET Lite Data. EDATA composed of ECHONET Object (EOJ) for both Source and Destination, ECHONET Lite Service (ESV), Number of processing properties (OPC), ECHONET Lite Property (EPC), Property data counter (PDC) and property value data (EDT). Both EHD1 and EHD2 will have 1 byte, where EHD1 is used to indicate the ECHONET Lite Protocol while EHD2 is used to determine the EDATA frame format. A value of 0x81 (hexadecimal, 10000001 in binary) in EHD2 to indicate the EDATA is in format 1 (specified message format) else 0x82 (hexadecimal, 10000010 in binary) to specified EDATA is in format 2 (arbitrary message format). TID is used to uniquely identify for each response that the sender received. Requests that do need to have a response, value for TID will not expressly specified. EOJ contained information about a particular object (electrical appliances, sensors and etc.)

3 Smart Home Using MEMS Sensor and Ambient Intelligence (SAHOMASI) Implementation

In this section will present the implementation of SAHOMASI.

3.1 Architecture of SAHOMASI System

Fig. 2 shows the overall system architecture of SAHOMASI system. Various sensors will be installed in the home which is attached to an Arduino board. Data is transmitted serially to Raspberry Pi for processing and data storing into the server. The mobile device is used for controlling home appliance setting and is used for fall detection. A notification or alert will be sent to the caregiver at certain situations whenever the system has detected abnormal activity or a fall occurred to the elderly.

The web application is used for various activity control and manage by both Caregiver and Admin. Both of them will have different privileges and has different actions that can be performed on the web application. Caregiver will use the web application to carry out modifying caregiver and elderly information, adding, editing, deleting and viewing for both sensor and location information and so on. While for admin they will be performing action such as approve for a request, view feedback made by caregiver and others.

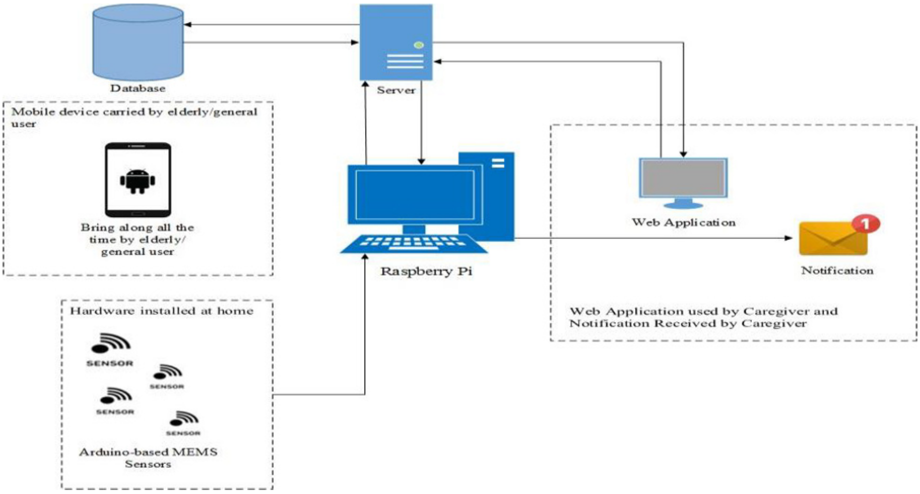

Database
Server
Mobile device carried by elderly/general user
Bring along all the time by elderly/ general user
Raspberry Pi
Web Application
Notification
Web Application used by Caregiver and Notification Received by Caregiver
Hardware installed at home
SENSOR
SENSOR
SENSOR
SENSOR
Arduino-based MEMS Sensors

Fig. 2. Overall System Architecture for SAHOMASI

3.2 Communication Protocol Architecture

Fig. 3 shows SAHOMASI system with ECHONET Lite as the standard communication protocol. Basically, the communication layer in ECHONET Lite contained of 3 layers, namely application software, ECHONET Lite communication middleware, and lower communication layer. Application software can be categorized into 2 types, which are software that provide control of devices and software that realizes the hardware function [10]. While for ECHONET Lite communication middleware is located in between of application software and lower communication layer that will process communication with respect to ECHONET Lite communication protocol [10]. For lower communication layer, it is responsible for data transmission and communication protocol processing that are unique to each transmission medium [10].

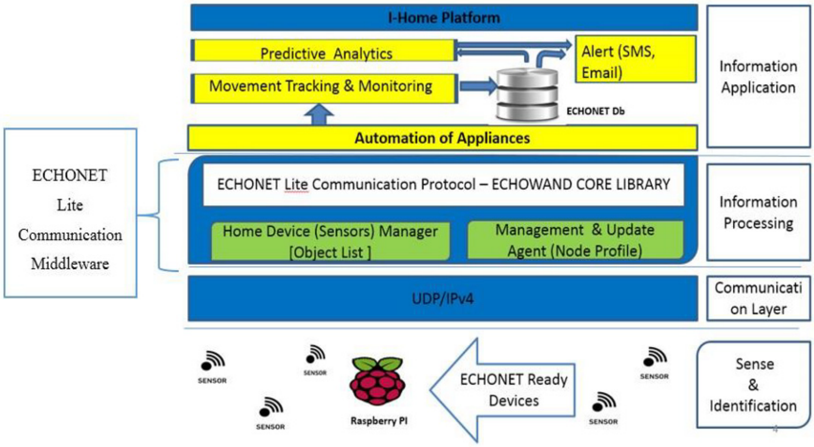


Fig. 3. Overall Communication Protocol Architecture for SAHOMSAI (ECHONET Lite)

First all different devices and different sensors with different communication protocol will be delegated and become an ECHONET Ready Devices. Then the devices and sensors are ready for sensing the activities and transmit data to ECHONET Lite communication middleware through Raspberry Pi. The ECHONET Lite communication middleware will convert the passed data packet from Raspberry Pi to its standardized data packet and use this data packet's information either to store in a database or for a prediction algorithm purpose.

3.3 SAHOMASI Hardware Requirements

Arduino-based MEMS Sensors. Arduino boards are used with a series of sensors to develop the sensing smart home environment. Sensors used in this system include motion sensor (HC SR501) and pressure sensor (Force Sensitivitive Resistor 0.5". All the sensors will be programmed using Arduino programming and will be powered by Arduino board itself. These sensors will be responsible for sensing functionality. Data will then be sent to Raspberry Pi board serially for further data processing. Refer to fig. 4.

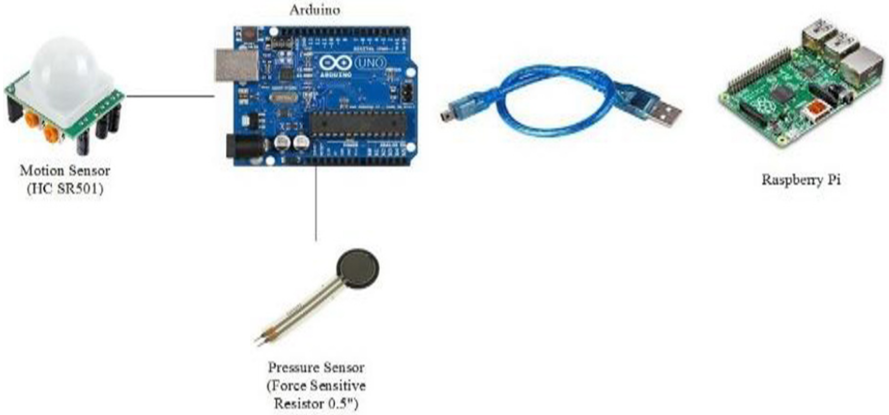


Fig. 4. Arduino-based MEMS Sensors connected serially with Raspberry Pi

Raspberry Pi-based Controller. Raspberry Pi is used as the core of the system. Raspberry Pi will receive data from Arduino board and store the data to the system server. Raspberry Pi will also be responsible for hosting the system intelligence and this intelligence includes abnormal activity detection, human movement activity prediction, as well as notification services.

Server. The server in this system will be used to host the database software which is myphpadmin. All data from Arduino-based sensors will be stored in these servers for future use by the system intelligences. Caregiver's information and elderly's information will also be stored in the same database. Meanwhile, the server will also be hosting the system's web application.

3.4 Fall Detection & Analytics Analysis & Algorithms Implementation

Fall Detection Algorithm. The system fall detection algorithm was adapted from the paper [12], where a smartphone's accelerometer is used in the implementation. Thus, fall detection services will be programmed in Android based operating system. The fall detection algorithm relies on accelerometer data from the built in accelerometer sensor for each smartphone.

Prediction Algorithm (Case-based Reasoning, CBR). The prediction algorithm is an algorithm that used in this system by which sensor data is analyzed to generate the schedule of daily activities of elderly. The generated daily activities contain the element of name of the activity, start time, end time and its location. The algorithm is as below:

- Retrieve sensor data for the past three days and store them in 3 lists according to the date and sort the data in each list according to the timestamp.
- For each list, partition the sensor data that come from same location and having timestamp difference of less than 5 minutes together with their sensor involved.

- For each partitioned sensor data is then sent to match best suited case from the case base, and determine the start time and end time.
- Merge the adapted cases from all three days into a single analysed activity.

3.5 Case Study of SAHOMASI based on Various Scenarios

Elderly experiences a fall and caregiver receives the notification. In this scenario, the elderly experiences a fall and is bringing the mobile phone that has the fall detection application on with them. The mobile application will continuously to monitor whether a fall has occur or not. If a fall occurred, the mobile application will log a data to the database, and and email will be sent to the caregiver immediately to inform the caregiver.

Realtime data monitor when elderly moving from one location to another. Cargiver can monitor the elderly movement activity through the realtime application of SAHOMASI system. When the elderly moving around in the home environment and triggered the sensor, sensor data will be updated in the application and a data will be logged to the database together with the timestamp of the occurance. The naming of each sensor will be named according the naming convention of ECHONET Lite. A 6 digits number will be used for the naming convention, the first 2 number will be used to represent the class group of ECHONET Lite, third number and fourth number will be used to represent what kinds of specific class is used (example, motion sensor will be represented by using digits “07”, and “10” for pressure sensor) and the last 2 digits is used as an instance identifier. A more detail of the naming convention n please refer to the ECHONET Lite Specification Appendix document.

Abnormal behavior detected and send notification to caregiver. The abnormal behavior of elderly includes fail or missed of carried out predicted activity. The system will continuously monitoring the elderly activity in the house, if the elderly missed one of the analysed acitivity, an email notification will be sent to caregiver to notify them that the elderly has failed to carriy out a specific analysed activity, so that the caregiver can take action immediately or to check on the elderly status. Fig. 8 shows a sample email that is sent to the caregiver when the elderly has failed to carry out a specific anlysed activity.

4. SAHOMASI System Discussion

In this section, the discussion on the prototype would be constructed in term of the physical layout, middleware structure, applications and analytics.

In our designed prototype, by employing MEMS sensors such as humidity, temperature and others, we have successfully converted each sensor to an Echonet object by understanding the behavior of input and output for each sensor. Some sensors are simply using *On* and *Off* as the output value; while others adopts the ranging values as the threshold parameters in triggering rules of output. The conversion of the Echonet involve the pre-planning especially in grouping sensors which has similar functions to be group together under similar IP address or subneting address. By employing Raspberry Pi devices to behave as a microprocessor for the prototype, each Raspberry Pi has been allocated unique address. The powerful middleware of Echonet allow the usage of third part microprocessor such as Raspberry and Beagle Bone Black to become an Echonet adapter. This supports the designed of a cost-saving implementation of Smart home prototype.

In term of the middleware structure, Echonet middleware with the advancement provided by Echonet light API which is programmed with Java has the property of completeness. These properties are important in designing a prototype which involves various kind of sensors. However, there is lack of classes for certain situation such as a user fall detection. By employing the human detection classes within Echonet, there were still issues in term of initializing the movement due to the need to initialize the human movement through the axis X,Y and Z with the support of accelerometer.

In this prototype, we have simply use mobile accelerometer and fail to initialize the sensor as Echonet adapter. The system will benefit to the caregiver and elderly by providing a better and caring home environment without the needs of full attentions from the caregiver. With the standardized communication protocol of this system, this system can be used with any appliances and sensors that come with different communication protocol. A communication between those appliances and sensors will be more easily and effortlessly. In future if more appliances and sensors are needed to be used or integrated into the system, this can be made easily without having to change or do configuration on those appliances or sensors that doesn't comply with the existing used communication protocol.

In this section, a SWOT analysis is conducted to give the overall picture of SAHOMASI system. The system is able to detect for a fall and sense for abnormal activities. When something fishy happened to the elderly, the system is able to alert the respective caregivers and any other recipients that the caregiver added to the system. In addition, having a standardized communication protocol, make the data exchanges and data communication between devices more easily and more effectively.

In contrast to the strength, the system has its limitation as well. The first limitation is that, a mobile phone based fall detection application is considered bulky to be carried around and is very fragile due to its touch screen specification. When the mobile phone is dropped or when the elderly fall down, there is a chance the screen will crack and cause the phone to be broken or spoiled. Thus, continuous replacing of broken mobile phone will be very costly. Besides, when the elderly forgot to bring along the mobile phone, then the fall detection application will be able to detect whether fall occurs to the elderly or not. Lastly MEMS sensors rely heavily on the internet to send data to the database, if there is a network services' failure, then data will not be able to store to the database and MEMS sensors used will not be able to differentiate if there is more than one person at the location. For instance, motion sensor in the living room can only detect the presence of people, but are not be able to differentiate the exact person that is under the system monitoring and protection.

5 Conclusion And Future Work

In a nutshell, this system is able to help elderly live independently by monitoring and tracking the elderly real-time. Meanwhile, reduce the burden and effort required for the caregiver to take care of the elderly in this modern world children of the elderly are busy with their works and the caregiver services by third party is in increasing trend.

As part of the future part, there are a couple ways to improve this system:

- Additional functionalities should be added to the system such as home automation and smart appliances controlling system using mobile phone or wearable device.
- A wearable device-based fall detection system should be implemented to replace the existing smart-phone based fall detection due to the limitation of using a smart phone.

Acknowledgement
This work is supported by a Research Grant provided by University Sains Malaysia. This work is a collaborative work between Japan Advanced Institute Sciences and Technology (JAIST and Collaborative Microelectronic Design Excellence Centre (CEDEC), University Sains Malaysia

References

1. Sripan, M., et al. Research and thinking of smart home technology. in International Conference on Systems and Electronic Engineering-(ICSEE'2012. 2012.
2. Poland, M.P., et al., Smart home research: Projects and issues. International Journal of Ambient Computing and Intelligence (IJACI), 2009. **1**(4): p. 32-45.
3. Killeen, P., et al. Developing a Smart Home System. in Proceedings of the International Conference on Embedded Systems and Applications (ESA). 2011. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
4. Lin, H.-T., Implementing smart homes with open source solutions. International Journal of Smart Home, 2013. **7**(4): p. 289-295.
5. Piyare, R. and S.R. Lee, Smart home-control and monitoring system using smart phone. ICCA, ASTL, 2013. **24**: p. 83-86.
6. Cook, D.J., J.C. Augusto, and V.R. Jakkula, Ambient intelligence: Technologies, applications, and opportunities. Pervasive and Mobile Computing, 2009. **5**(4): p. 277-298.
7. Cook, D.J., et al. MavHome: An agent-based smart home. in Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on. 2003. IEEE.
8. Kit, J.L.W., M.M. Singh, and N.H.A.H. Malim, Ambient Intelligence Smart Home Automation (AMISHA) System. Advanced Science Letters, 2017. **23**(6): p. 5073-5077.
9. Lim Poon Koon and M.M. Singh, iHOME: An Ambient Intelligence Mobile CrowdSensing Smart Home System, in Knowledge Management International Conference (KMICe) 2016. 2016: Chiang Mai, Thailand.
10. Consortium, E. ECHONET Lite Specification ver1.12 Part 1: ECHONET Lite Overview. 2016.

11. Consortium, E. ECHONET Lite Specification ver1.12 Part 2: ECHONET Lite Overview. 2016.
12. Cao, Y., Y. Yang, and W. Liu. E-FallD: A fall detection system using android-based smartphone. in Fuzzy systems and knowledge discovery (FSKD), 2012 9th international conference on. 2012. IEEE.

Context Aware Knowledge Bases for Efficient Contextual Retrieval: Design and Methodologies

Sharyar Wani[1], Tengku Mohd. Tengku Sembok[2] and Mohammad Shuaib Mir[3]

[1] Cyber Security Centre, National Defence University of Malaysia, Kuala Lumpur, Malaysia
[2] Faculty of Defence Science and Technology, National Defence University of Malaysia, Kuala Lumpur, Malaysia
[3] Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia
sharyar@upnm.edu.my,tmts@upnm.edu.my,shuaib.mir@live.iium.edu.my

Abstract. Contextual retrieval is a critical component for efficient usage of knowledge hidden behind the data. It is also among the most important factors for user satisfaction. It essentially comprise of two equally important parts – the retrieval mechanism and the knowledge base from which the information is retrieved. Despite the importance, context aware knowledge bases have not received much attention and thereby, limiting the efficiency of precise context aware retrieval. Such knowledge bases would not only contain information that has been efficiently stored but the knowledge contained would be context based. In other words, machines would understand the knowledge and its context rather than just storing data. This would help in efficient and context aware retrieval. The current paper proposes rules and methodologies for construction of such context aware knowledge bases. A case study to demonstrate the application of the methodology and test the efficiency of the proposed methodology has also been presented. The results indicate that knowledge bases built on these principles tend to generate more efficient and better context aware retrieval results.

Keywords: Contextual Storage, Contextual Knowledge Base, Context Aware Knowledge Base, Contextual Retrieval, Contextual Understanding, Contextual Machine Understanding.

1 Introduction

Contextual information retrieval stands out as a major problem, even after decade's long research in information retrieval. This problem has gained more importance after the advances and popularity in semantic web, natural language processing and other areas of artificial intelligence. Users tend to be highly dissatisfied if the first few retrieval results aren't the correct answers to their search terms/questions. The dissatisfaction increases further, if the retrieved results/answers are non-contextual. Traditionally, contextual retrieval is understood as combination of search mechanisms, query understanding, and user context, operating in a single framework. Therefore,

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_55

the understanding implies that the issues of contextual retrieval are only linked to the retrieval side. However, the results achieved so far aren't satisfactory enough. Therefore, it is important to recognize that the aforementioned understanding only contributes to half of the process. The other half of the process is the construction of knowledge bases which store in context. In other words, machines need to understand the extracted knowledge in context. While a lot of research has been directed towards the retrieval side, not much has been done in regards to construction of more meaningful knowledge bases aka knowledge bases in context or extracting, representing and storing knowledge in context.

What does a search engine or a question answer system understand when a user searches for "best laptop"? The retrieval systems could suggest lists of outstanding laptops that exist in the market. These lists are usually based on ratings given by different agencies, technology portals or third party review services, etc. However, these search systems don't consider producing custom results from the list of user reviewed laptops and display the results based on categories such as best processing laptops, best gaming laptops, best laptops according based on data transfer, etc. One might question that the user should have specifically asked for such categories. However, users tend not to write specific queries most of the time. Not that they don't know how to write an intelligent query but because of other reasons such as non tech savvy people don't know or remember the terminologies beforehand or they want to see diverse results before they choose. Although it may seem more a grouping problem, it is the context for the user in this example. The current results make it cumbersome as the users have to scroll through all the pages and compare different results from numerous pages before reaching a conclusion. If one was looking for high processing power, he/she might have to open multiple pages and chose the high processing machines from each of them and perform a self-comparison. Better retrieval precision is required in such a case. The retrieval precision is more so required in critical domains such as health etc.

It is important to understand that many a times the users won't be specific in their search terms or might pose vague questions. In such cases, the user intent, which is an important part of an effective contextual retrieval, is missing and thus reducing the chances of retrieving precise information. However, this limitation can also be overcome by designing context aware knowledge bases. Secondly, the above example also demonstrates that the set of results won't be completely relevant since the user query and the intent within have a large scope. Therefore, the search system needs to take into account the process of filtering irrelevant results. It should also be noted that users also tend to submit incomplete search terms/queries, thereby, increasing the work load of the system to produce the contextually precise results.

Although the retrieval part of these systems have an important role to play to solve the aforementioned problems but they cannot deliver efficient results without considering the knowledge repositories they utilize. Therefore, it is important that machines understand the knowledge behind the data as well as the context of that knowledge.

Although a lot of work has been done in improving the search mechanisms, personalized search and contextual search, to our best knowledge, not much work has been directed towards context aware knowledge bases. While improvements have been made on the retrieval part, the knowledge representation has been left out of the process. This paper focuses on methodologies to build context aware knowledge bases in a systematic way which will address the key issues and provide solutions for precise and efficient context aware retrieval.

The current paper discusses the novel concept of using semantic graphs with logic and linguistics as its core drivers. The formulated methodology breaks down the data to its atomic level while maintaining the relations between the data/text. The logic is used to determine, maintain and index the entities inside these graphs while as linguistics determine the pattern of logical arrangement, context of sentences/words, relations between words/sentences, and semantic analysis. This is applicable both within a document and between other documents. Based on these concepts, we propose rules and methodologies to construct efficient context aware knowledge bases. We also conduct a case study to demonstrate the entire process and its context aware efficiency. Our experimental results demonstrate that the proposed rules and methodologies can achieve significantly better context aware retrieval performance. The same concept can be extended to other data domains.

The rest of the paper is organized as follows: Section 2 presents the details of the methodology. Section 3 discusses the rules employed for designing context aware knowledge bases. The experimental results have been discussed further in Section 4. It is followed by closely related work, conclusion and future work in Sections 5 and 6 respectively.

2 Related Work

Relation extraction helps in determining the type of relationship between two target entities that appear together in a text. In a context aware knowledge base, this is of great value. Demonstrate that it is beneficial to consider multiple relations in the sentential context when predicting target relation between two entities. The researchers use an LTSM based encoder to learn all the relations from a particular sentence. The target relation and context relations are combined to make the final prediction. However, this methodology preserves the final prediction as the only relation, neglecting other relations which hold equally important value in the sentence, paragraph, documents, etc. [3]

Similarly Hoffman present multi instance learning handling overlapping relations at a sentence level. The work focuses on binary level relations involving nouns. It is based on the idea that a fact may be true for multiple entities such that r and q hold true for both $r(e1,e2)$ and $q(e1,e2)$ [4].

Reference [5] uses context aware methodology in knowledge base completion. They propose C-PR, an algorithm that learns global semantics of entities in a knowledge base using word embedding approach. Then, it uses the entity semantics information to compute contextually relevant paths for knowledge base completion.

These are the closest related works, to the best of our knowledge, who have considered the concept of context aware knowledge bases whether directly or indirectly. However, a lot of work has been done on the retrieval side for context aware retrieval e.g. retrieval based on spatial context [6], probabilistic model for contextual retrieval [7] [8], correspondence analysis [9], contextual image retrieval [10], etc.

3 Methodology

In order for a search engine or question answer system or any other expert system to retrieve efficient and precise contextual results, both the knowledge base and the retrieval methodology are equally important. In a more efficient system, they are actually interdependent on each other based on the design of their methodologies. As discussed earlier, retrieval methodologies have received some focus comparatively than its counterpart. However, efficient precise contextual results are impossible without improving both the sides of a particular system. Therefore, this paper focuses on the rules, principles and methodologies for construction of context aware knowledge bases.

Before delving further in discussing the mechanisms, it is important to explain the statement that in highly efficient systems, the principles/design of construction of knowledge bases and retrieval strategies are same. This can be explained simply using the following logic; if a knowledge base ***K*** was based on the design pattern

$$k = s1 + s2 + s3 \tag{1}$$
$$s1 = n1 + n2 + n3 + n4 \tag{2}$$
$$s2 = n4 + n5 + n6 + n7 \tag{3}$$
$$s3 = n8 + n9 + n10 + n2 \tag{4}$$

where ***s*** indicate the sentences inside a document and ***n*** are the nodes inside a graph database storing the construction of these sentences. Based on this, an effective retrieval strategy should be designed using the same design in a backward direction. This implies that a search for node n3 from a query ***q*** would be processed and searched as follows

$$q = n12 + n13 + n14 + n15 \tag{5}$$

where n13 is a semantic equivalent of n3 and the required search field. Therefore, when the query is broken down or designed in the same manner as the knowledge base was designed, semantic matching amongst the nodes from the two sides is easier, effective and efficient.

Our design for context aware knowledge bases starts with the thorough linguistic analysis of the document text. The analysis is performed and a semantic tag is associated with each component of the document structure. A second classification is performed to determine and categorize the most essential components of the document/sentences that are necessary to determine the context of each component of the sentence. The non-essential or more accurately least essential are categorized separately for later stages to enhance the semantic connectivity in the later phases of the design stage. This categorization is followed by restructuring/logical arrangement of the context aware component space aka phrases. The arrangement is based on first order predicate logic. This helps to establish a consistent pattern of arrangement for all sentences of similar structure. The pattern is further re-shaped into a higher level arrangement where the relations are purely based on the connections and relatedness using the predicate subject principles of first order calculus rather than semantic based logical arrangement of the previous step. Therefore, two different yet associated patterns are designed for each text structure. Based on this, semantic networks are established inside and between the documents to complete the construction of the knowledge base. The semantic networks act as culmination or a framework for different types of sentence structures helping in extraction, representation and storage of context aware knowledge.

4 Context Aware Design

Documents comprise of multiple number of sentences forming passages and paragraphs. A paragraph often contains a contextual discussion compared to a passage. The elements of a paragraph are interlinked to each other until the end of discussion in order to complete the context of the underlying discussion. On the other hand, a passage might not possess all the qualities of a contextual discussion. A passage sometimes jumps to a slightly new discussion, although it might be related. This can be referred to as the second degree contextual element. Based on this, a paragraph still lies under the same contextual discussion which might be referred as the first degree context for ease of discussion. Consider the following passage ***P*** (Table 1) about the renowned scientist Marie Curie:

Table 1. Sample text

> *"Marie was born in 1867 in Warsaw, Poland, where her father was a professor of physics. At an early age, she displayed a brilliant mind and a blithe personality. Her great exuberance for learning prompted her to continue with her studies after high school. She became disgruntled, however, when she learned that the university in Warsaw was closed to women. Determined to receive a higher education, she defiantly left Poland and in 1891 entered the Sorbonne, a French university, where she earned her master's degree and doctorate in physics.* **(P1)**
> *Marie was fortunate to have studied at the Sorbonne with some of the greatest scientists of her day, one of whom was Pierre Curie. Marie and Pierre were married in 1895 and spent many productive years working together in the physics laboratory. A short time after they discovered radium, Pierre was killed by a horse-drawn wagon in 1906. Marie was stunned by this horrible misfortune and endured heartbreaking anguish. Despondently she recalled their close relationship and the joy that they had shared in scientific research. The fact that she had two young daughters to raise by herself greatly increased her distress."* **(P2)**

The above passage consists of two paragraphs. Paragraph ***P1*** focuses on the background of her personality in regards to her education. ***P2*** essentially discusses her family life. Although the passage discussed different details about the scientist's life, each paragraph focuses on a particular context and the discussion is concluded by the end of the paragraph so that the context is clear, concise and well understood.

Further, paragraphs consist of sentences and sentences are construction of words/phrases. A sentence represents a part of discussion within a context. Every sentence essentially has a context of its own which relates and is in accordance with the context of the paragraph. Phrases are the fundamental units of construction of sentences which form sentences, paragraphs, passages, stories, etc. The most essential phrases within a sentence are the verb phrases and noun phrases. Verbs denote action, state, or occurrence and form the main part of the predicate of the sentence. Nouns signify the presence of class, people or things. In the current context, nouns describe the doer of the verb. It is the combination of these nouns and verbs that essentially signify the context of the speech. However, it should be noted that the remaining components of the sentence are important as well. These components help to maintain the relationships between the nouns and verbs in formal English. In context of the current work, these components are used to enhance semantic relationships rather than establish it as in case of formal English.

Consider one of the sentences from the above passage e.g.; "*She became disgruntled, however, when she learned that the university in War-saw was closed to women.*" A constituency parse would give the following output:

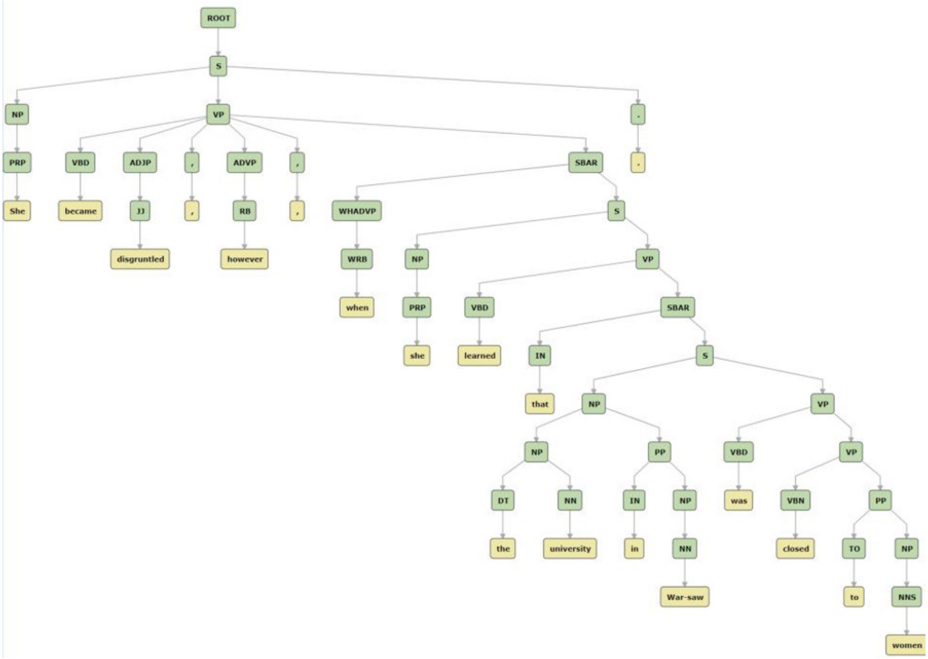


Fig. 1. Constituency parse tree of the sample sentence.

A typical Information Extraction model for the same would be as follows shown in Fig. 2.

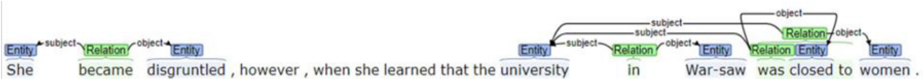


Fig. 2. Information extraction of the sample sentence

A quick analysis of the constituency parse and information extraction model reveals that certain phrases which are actually a part of the verb phrase and noun phrase have been omitted while defining the relationships or formation of tuples of information in information extraction process. While this is correct when measuring the output against formal English, however, a lot of information and context is lost while eliminating such phrases. Modern approaches focus only on the immediate relations between entities disregarding other potential relations present in the same sentence [1-3]. Consider the phrase "learned" which is a part of the verb phrase and has been omitted in information extraction model. Firstly, the word describes the context of the information preceeding it. Secondly, if a user were to inquire "Why did Marie feel disgruntled?", the context and relationship between being disgruntles and its reason is lost. This is a small example of the absence of context. We propose that all verb phrases and noun phrases should be present when presenting information inside a knowledge base and should be connected to each other in terms of their semantic relatdness. Secondly, we propose that these phrases should be re-organized using the first order predicate logic where the first verb phrase represents the predicate, the following noun phrase represents the subject. Further, the noun phrases or verb phrases or sets of their combinatins represent objects for the sentence. The reason beind using the first verb as the predicate is based on the notion that the first verb phrase provides the context to the whole sentence. The accompanying noun phrase is typically the entity under discussion related to the verb. The remaining phrases are further details to the main discussion, hence the objects. In other words, each phrase helps to maintain the context of discussion which is primarily defined by the first verb phrase and noun phrase of the sentence. Based on this methodology, the knowledge extraction would be represented as follows:

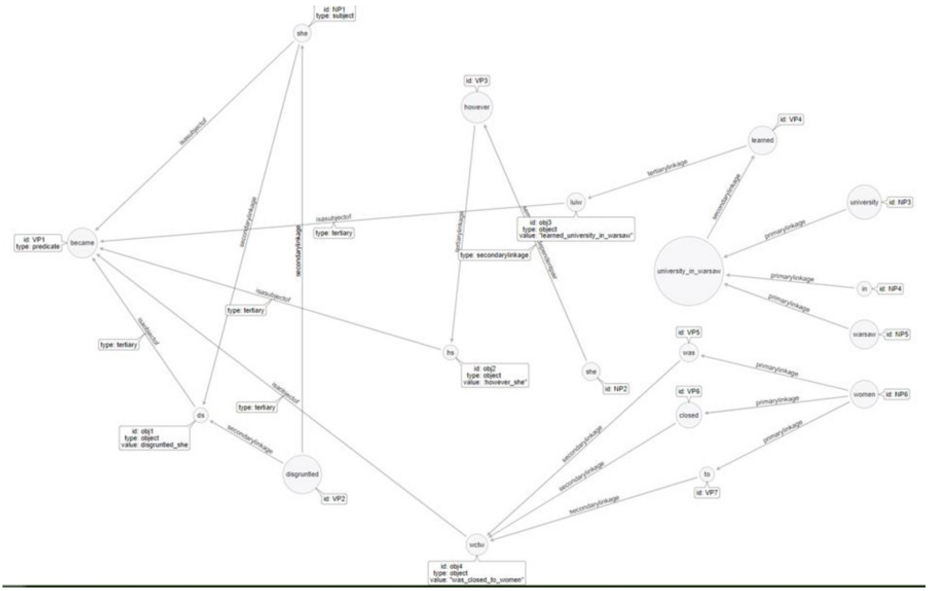


Fig. 3. Context aware knowledge base for the sample sentence

The cypher representation for the above graph is as follows:

Table 2. Cypher Representation of the context aware knowledge base.

```
(`0` :became {id:'VP1',type:'predicate'}) ,
(`1` :she {id:'NP1',type:'subject'}) ,
(`2` :disgruntled {id:'VP2'}) ,
(`3` :however {id:'VP3'}) ,
(`4` :she {id:'NP2'}) ,
(`5` :hs {id:'obj2',type:'object',value:':however_she"'}) ,
(`6` :university_in_warsaw ) ,
(`7` :university {id:'NP3'}) ,
(`8` :in {id:'NP4'}) ,
(`9` :warsaw {id:'NP5'}) ,
(`10` :was {id:'VP5'}) ,
(`11` :learned {id:'VP4'}) ,
(`12` :ds {id:'obj1',type:'object',value:'disgruntled_she'}) ,
(`13` :closed {id:'VP6'}) ,
(`14` :to {id:'VP7'}) ,
(`15` :women {id:'NP6'}) ,
(`16` :luiw {id:'obj3',type:'object',value:"learned_university_in_warsaw"}) ,
(`17` :wctw {id:'obj4',type:'object',value:"was_closed_to_women"}) ,
(`1`)-[:`isasubjectof` ]->(`0`),
(`2`)-[:`secondarylinkage` ]->(`1`),
(`4`)-[:`semanticallydependentpair` {type:'secondarylinkage'}]->(`3`),
(`3`)-[:`tertiarylinkage` ]->(`5`),
(`5`)-[:`isasubjectof` {type:'tertiary'}]->(`0`),
(`7`)-[:`primarylinkage` ]->(`6`),
(`8`)-[:`primarylinkage` ]->(`6`),
(`9`)-[:`primarylinkage` ]->(`6`),
(`6`)-[:`secondarylinkage` ]->(`11`),
(`2`)-[:`secondarylinkage` ]->(`12`),
(`1`)-[:`secondarylinkage` ]->(`12`),
(`12`)-[:`isaobjectof` {type:'tertiary'}]->(`0`),
(`11`)-[:`tertiarylinkage` ]->(`16`),
(`16`)-[:`isasubjectof` {type:'tertiary'}]->(`0`),
(`15`)-[:`primarylinkage` ]->(`10`),
(`15`)-[:`primarylinkage` ]->(`13`),
(`15`)-[:`primarylinkage` ]->(`14`),
(`10`)-[:`secondarylinkage` ]->(`17`),
(`13`)-[:`secondarylinkage` ]->(`17`),
(`14`)-[:`secondarylinkage` ]->(`17`),
(`17`)-[:`isaobjectof` {type:'tertiary'}]->(`0`)
```

Hence, logic is being used to determine, maintain and index the entities inside these graphs while as linguistics determine the pattern of logical arrangement, context of phrases, relations between them and semantic analysis as a whole. This is a context aware model and methodology at the basic atomic level i.e.; phrases, in this case verb phrases and noun phrases. Further, the context awareness will be enhanced at word level, paragraph level, passage level, etc. which will be presented in our future work.

5 Experimentation

5.1 Data Set

The dataset for the experiment consists of over 5000 documents in text format. These documents were converted into the context aware knowledge base design as presented above. We tested the context aware effectiveness using 28 different questions. Each question was presented in two different styles to accommodate the diverse thinking of users to a certain extent. Therefore, the system was tested against a total of 56 questions.

5.2 Experimental Results

This section provides the empirical evidence about the effectiveness of our contextual aware databases for context aware retrieval. FHS, FARR, TRR were used as units of measurement.

Table 3. Experimental results

Metric	Average Value
First Hit Success (FHS)	0.75
First Answer Reciprocal Rank (FARR)	0.83
Total Reciprocal Rank (TRR)	0.87

FHS indicates that the system is capable of answering contextually, 8 out of 10 times as the first answer. FARR indicates that the user's effort of finding the correct answer is reduced by 83% in the list of answers. This implies that only 1 out of 10 of the results bear less or no context awareness to the query. TRR is indicative that almost 87% answers are related to the context of the question. In other words, the set of results is 90% contextually related or context aware.

6 Conclusion

The paper discussed a new methodology for construction of context aware knowledge bases. This helps to solve some of the problems of contextual information retrieval. While has a lot of work has been dedicated towards the retrieval part of contextual retrieval, not much attention has been paid to the knowledge base side. This work focuses in that domain so that the efficiency of contextual retrieval can be maximized. The experimental results are indicative that the new methodology is efficient and sound in nature and greatly increases the chances of obtaining precise contextual results.

The paper focuses on usage of logic, linguistics and semantics knowledge networks coherently supported by linking relationship methodologies and rules to design knowledge bases where relations are observed to the minutest level. Thereby, this will help to design more efficient and meaningful context aware knowledge bases. The current paper demonstrated the context aware methodology at the phrase level inside the sentences. In future, we will present context aware methodologies at word, paragraph, passage, prose, etc. aka higher construction levels of a language/data.

References

1. Zeng, D., et al. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.
2. Lin, Y., et al. Neural relation extraction with selective attention over instances. in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.
3. Sorokin, D. and I. Gurevych. Context-Aware Representations for Knowledge Base Relation Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.
4. Hoffmann, R., et al. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for

Computational Linguistics: Human Language Technologies-Volume 1. 2011. Association for Computational Linguistics.

5. Mazumder, S. and B. Liu, Context-aware path ranking for knowledge base completion. arXiv preprint arXiv:1712.07745, 2017.
6. Palacio, D., et al., Prototyping a personalized contextual retrieval framework, in Proceedings of the 7th Workshop on Geographic Information Retrieval. 2013, ACM: Orlando, Florida. p. 43-44.
7. Wen, J.-R., N. Lao, and W.-Y. Ma, Probabilistic model for contextual retrieval, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004, ACM: Sheffield, United Kingdom. p. 57-63.
8. Huang, X., et al., A platform for Okapi-based contextual information retrieval, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006, ACM: Seattle, Washington, USA. p. 728-728.
9. Freund, L., E.G. Toms, and C.L.A. Clarke, Modeling task-genre relationships for IR in the workplace, in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005, ACM: Salvador, Brazil. p. 441-448.
10. Xing, X., Y. Zhang, and B. Gong, Mixture model based contextual image retrieval, in Proceedings of the ACM International Conference on Image and Video Retrieval. 2010, ACM: Xi'an, China. p. 251-258.

Correction to: A Review on Energy Efficient Path Planning Algorithms for Unmanned Air Vehicles

Sanjoy Kumar Debnath, Rosli Omar and Nor Badariyah Abdul Latip

Correction to:
Chapter "A Review on Energy Efficient Path Planning Algorithms for Unmanned Air Vehicles" in: R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_51

The original version of this chapter was inadvertently published with incorrect first author's name as "Sulaiman Sanjoy Kumar Debnath". This has been corrected as "Sanjoy Kumar Debnath" in the chapter.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-981-13-2622-6_51

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes
in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6_56

Author Index

A
Abdalla, Mansour Hassani, 67, 145
Adnan, Ramli bin, 269
Agarwal, Abhishek, 155
Ahmad, Rabiah, 513
Ahmad, Waseem, 369, 461
Alabsi, Basim, 347
Al-Ani, Ahmed K., 305, 315, 347
Al-Ani, Ayman, 305, 315
Alfred, Rayner, 67, 145, 215, 227
Al-Ghaili, Abbas M., 107
Ambar, Radzi, 533
Anbar, Mohammed, 305, 315, 347
Anthony, Patricia, 471, 481
Anuar, Nor Badrul, 59
Arshad, Nor Hashim Mohd, 269
Aziz, Nasim, 513
Azmi, Tengku Muhammad Afif bin Tengku, 493

B
Baharum, Aslina, 87, 533
Belaton, Bahari, 347
Bolongkikit, Jetol, 67, 145, 175, 215, 227
Bono, A., 97, 359
Bulagang, Aaron Frederick, 87

C
Chew, I. M., 97, 359
Chiewchan, Kitti, 471
Chin, Zi Hau, 117
Chua, Sook-Ling, 237, 249
Chuan, Bernard Lee Jin, 545
Connie, Tee, 25, 37

D
Dadiz, Bryan G., 413
Das, Joydeep, 155
Das, Soumita, 155
Dasril, Yosza, 401
Debnath, Sanjoy Kumar, 523

E
Elejla, Omar E., 347
Eng, Jeng Hong, 127

F
FANG, Yuan, 165
Fang, Yuan, 185
Fauzi, Mohammad Faizal Ahmad, 1
Feizollah, Ali, 59
Foo, Lee Kien, 237, 249

G
Gabda, Darmesah, 401
Ghazali, Khadizah, 401
Ghita, Bogdan, 295
Goh, Kah Ong Michael, 37
Goh, Vik Tor, 117, 137, 259, 445

H
Hamdan, Abdul Razak, 481
Haque, Muhammad Reazul, 195
Hassan, Fadratul Hafinaz, 283
Hassan, Zainuddin, 107
Haw, Su Cheng, 1
Hipiny, Irwandi Hipni Mohamad, 77
Ho, Chin-Kuan, 249
Ho, Chiung Ching, 117, 137
Hoe, Lau Siong, 25

R. Alfred et al. (eds.), *Computational Science and Technology*, Lecture Notes in Electrical Engineering 481, https://doi.org/10.1007/978-981-13-2622-6

Htun, Naing Linn, 423

I
Ibrahim, Najihah, 283
Ibrahim, Zul-Azri, 107
Islam, Amanul, 503

J
Jali, Shaiful Hisham Mat, 13
Jhinn, Wee Lorn, 25
Juboor, Saed Sa'deh, 237
June, Tan Li, 175
Jung, Euihyun, 453
Junn, Kuan Yik, 215, 227

K
Kasim, Hairoladenan, 107
Kaspin, Ri-zaludin, 195
Kassim, Mohamad Nizam, 13
Khambari, Najwan, 295
Kiah, Laiha Binti Mat, 59
Kim, Hyoungshick, 49
Koh, Jia Juang, 137
Kok, Leonard, 249
Kok, Woon Chee, 335
Ku, Chin Soon, 503
Kuek, Thiam Yong, 137
Kumar, Sudhindra, 59

L
Labadin, Jane, 335
Latip, Nor Badariyah Abdul, 523
Leau, Yu-Beng, 175, 305
Lee, Ching K., 195
LIM, Yuto, 165
Lim, Yuto, 185, 557

M
Maarof, Mohd Aizaini, 13
Majumder, Subhashis, 155
Mamat, Mazlina, 205
Manaf, Asrulnizam, 557
Manickam, Selvakumar, 305
Maseri, Marlyn, 205
Memon, Shuaib K., 369, 461
Minoi, Jacey-Lynn, 77
Mir, Mohammad Shuaib, 569
Moon, Jongho, 49
Mun, Ng Kok, 269

N
Nandong, J., 97, 359
Ng, Hu, 1, 117, 137, 259, 445
Ng, Jin Ren, 259
Nisar, Kashif, 369, 461

O
Obit, Joe H., 175
Obit, Joe Henry, 67, 145, 215, 227
Omar, Rosli, 523
On, Chin Kim, 481
Ong, Michael Goh Kah, 25
Ong, Thian Song, 37
OOI, Sian En, 165
Ooi, Sian En, 185
Othman, Fazidah, 503
Othman, Marini, 107

P
Peter, Marcella, 77
Phing, Tan Ching, 533
Poad, Hazwaj Mhd, 533
Por, Lip Yee, 503

R
Rahim, Fiza Abdul, 107
Ramly, Mohd Rizal Bin Mohd, 249
Ramos, Christine Diane, 379
Razak, Noorfadzli Abdul, 269
Ruiz, Conrado R., 413
Ryu, Jihyeon, 49

S
Samarasinghe, Sandhya, 471
San, Cho Cho, 423
Saudi, Azali, 127
Shah, Syafiq Bin Ibrahim, 315
Shariff, Azizul Rahman Mohd, 545
Sheng, Ong Yan, 227
Sing, Angeline Lee Ling, 391
Singh, Gurpreet, 461
Singh, Manmeet Mahinderjit, 545, 557
Song, Taeui, 49
Soon, Gan Kim, 481
Srikanth, Karanam, 59
Suarez, Merlin Teodosia, 379
Sulaiman, Jumat, 127, 401
Sulaiman, Nadzril bin, 493

T
Tai, Kain Hoe, 445
Tan, Saw C., 195
TAN, Yasuo, 165
Tan, Yasuo, 185
Tay, Nian Chi, 37
Teh, Pin Shen, 37
Tengku Sembok, Tengku Mohd., 569
Thamrin, Norashikin M., 269

Thinn, Aye Aye, 435
Thwin, Mie Mie Su, 423, 435
Tighe, Edward, 379
Tong, Hau Lee, 1, 117, 137

W
Wahab, Mohd Helmy Abd, 533
Wani, Sharyar, 569
Weng, Ng Giap, 391
Won, Dongho, 49
Wong, F., 97, 359
Wong, Jan Shao, 259
Wong, K. I., 97, 359

Y
Yap, Timothy Tzen Vun, 1, 117, 137, 259, 445
Yap, Wen Jiun, 259
Yunos, Zahri, 513
Yusoff, Zulfadzli, 195

Z
Zainal, Anazida, 13
Zakaria, Safial Aqbar, 283
Zaki, Fatimah Audah Md., 325
Zamanian, Zahedeh, 59
Zulkurnain, Nurul Fariza, 325