Lecture Notes in Electrical Engineering 514

Kuinam J. Kim · Nakhoon Baek
Editors

Information Science and Applications 2018

ICISA 2018

Springer

Lecture Notes in Electrical Engineering

Volume 514

** **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:

Asia:

China, *Jessie Guo, Assistant Editor* (jessie.guo@springer.com) (Engineering)

India, *Swati Meherishi, Senior Editor* (swati.meherishi@springer.com) (Engineering)

Japan, *Takeyuki Yonezawa, Editorial Director* (takeyuki.yonezawa@springer.com)
(Physical Sciences & Engineering)

South Korea, *Smith (Ahram) Chae, Associate Editor* (smith.chae@springer.com)
(Physical Sciences & Engineering)

Southeast Asia, *Ramesh Premnath, Editor* (ramesh.premnath@springer.com)
(Electrical Engineering)

South Asia, *Aninda Bose, Editor* (aninda.bose@springer.com) (Electrical Engineering)

Europe:

Leontina Di Cecco, Editor (Leontina.dicecco@springer.com)

(Applied Sciences and Engineering; Bio-Inspired Robotics, Medical Robotics, Bioengineering; Computational Methods & Models in Science, Medicine and Technology; Soft Computing; Philosophy of Modern Science and Technologies; Mechanical Engineering; Ocean and Naval Engineering; Water Management & Technology)

(christoph.baumann@springer.com)

(Heat and Mass Transfer, Signal Processing and Telecommunications, and Solid and Fluid Mechanics, and Engineering Materials)

North America:

Michael Luby, Editor (michael.luby@springer.com) (Mechanics; Materials)

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

Kuinam J. Kim · Nakhoon Baek
Editors

Information Science and Applications 2018

ICISA 2018

Springer

Editors
Kuinam J. Kim
iCatse
Seongnam, Gyeonggi
Korea (Republic of)

and

Kyonggi University
Suwon, Gyeonggi
Korea (Republic of)

Nakhoon Baek
School of Computer Science
and Engineering
Kyungpook National University
Daegu
Korea (Republic of)

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-13-1055-3 ISBN 978-981-13-1056-0 (eBook)
https://doi.org/10.1007/978-981-13-1056-0

Library of Congress Control Number: 2018945097

Printed on acid-free paper

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

This LNEE volume contains the papers presented at the iCatse International Conference on Information Science and Applications (ICISA 2018) which was held in Hong Kong, China, during June 25–27th, 2018.

ICISA2018 will provide an excellent international conference for sharing knowledge and results in Information Science and Application. The aim of the conference is to provide a platform to the researchers and practitioners from both academia as well as industry to meet the share cutting-edge development in the field.

The primary goal of the conference is to exchange, share, and distribute the latest research and theories from our international community. The conference will be held every year to make it an ideal platform for people to share views and experiences in Information Science and Application related fields.

On behalf of the Organizing Committee, we would like to thank Springer for publishing the proceedings of ICISA2018. We also would like to express our gratitude to the “Program Committee and Reviewers” for providing extra help in the review process. The quality of a refereed volume depends mainly on the expertise and dedication of the reviewers. We are indebted to the Program Committee members for their guidance and coordination in organizing the review process, and to the authors for contributing their research results to the conference.

Our sincere thanks to the Institute of Creative Advanced Technology, Engineering and Science for designing the conference web page and also spending countless days in preparing the final program in time for printing. We would also like to thank our organization committee for their hard work in sorting our manuscripts from our authors.

We look forward to seeing all of you next year at ICISA.

Suwon, Korea (Republic of) Kuinam J. Kim
Daegu, Korea (Republic of) Nakhoon Baek

Organizing Committee

General Chair

Nakhoon Baek, Kyungpook National University, Republic of Korea

Steering Committee

Nikolai Joukov, New York University and modelizeIT Inc, USA
Borko Furht, Florida Atlantic University, USA
Bezalel Gavish, Southern Methodist University, USA
Kin Fun Li, University of Victoria, Canada
Kuinam J. Kim, Kyonggi University, Korea
Naruemon Wattanapongsakorn, King Mongkut's University of Technology Thonburi, Thailand
Xiaoxia Huang, University of Science and Technology Beijing, China
Dong-Seong (Dan) Kim, University of Canterbury, New Zealand
Nakhoon Baek, Kyungpook National University, Republic of Korea

Publicity Chair

Hongseok Jeon, ETRI, Republic of Korea
Tomas Cerny, Czech Technical University, Czech Republic
Naruemon Wattanapongsakorn, King Mongkut's University of Technology Thonburi, Thailand
Suresh Thanakodi, National Defence University of Malaysia, Malaysia

Workshop Chair

Dong-Seong (Dan) Kim, University of Canterbury, New Zealand

Publication Chair

Kyoungho Choi, Institute of Creative Advanced Technologies, Science and Engineering

Program Chair

Kuinam J. Kim, Kyonggi University, Republic of Korea

Financial Chair

WonHyung Park, Institute of Creative Advanced Technologies, Science and Engineering
Sanggyoon Oh, BPU Holdings Corp, Republic of Korea

Organizers and Supporters

Institute of Creative Advanced Technologies, Science and Engineering (iCatse)
River Publishers, Netherlands
Korean Industry Security Forum (KISF)
Korea Convergence Security Association (KCSA)
Kyonggi University, Korea
King Mongkut's University of Technology Thonburi, Thailand
National Defence University of Malaysia, Malaysia
University of Canterbury, New Zealand
University of Science and Technology Beijing, China
Electronics and Telecommunications Research Institute (ETRI)
Korea Institute of Science and Technology Information (KISTI)
Kyungpook National University, Republic of Korea

Program Committee

Maicon Stihler, Federal Center for Technological Education of Minas Gerais—CEFETMG, Brazil
Zeeshan Ali Rana, FAST-NUCES, Lahore, Pakistan
Tan Syh Yuan, Multimedia University, Malaysia

Cimato Stelvio, Università degli Studi di Milano, Italy
Sira Yongchareon, Auckland University of Technology, New Zealand
Sharmistha P. Chatterjee, Florida Atlantic University, Boca Raton, USA
Etibar Seyidzade, Baku Engineering University, Azerbaijan
MAINGUENAUD Michel, Institut National des Sciences Appliquées, Rouen, Fance
Alessandro Bianchi, Departement of Informatics, University of Bari, Italy
Sandro Leuchter, Mannheim University of Applied Sciences, Germany
Michalis Pavlidis, University of Brighton, UK
Bernd E. Wolfinger, University of Hamburg, Hamburg, Germany
Manik Sharma, DAV University, Jalandhar, India
Somlak Wannarumon Kielarova, Faculty of Engineering, Naresuan University, Thailand
Jeril Kuriakose, Manipal University Jaipur, India
Shimpei Matsumoto, Hiroshima Institute of Technology, Japan
Yadigar Imamverdiyev, Institute of Information Technology of Azerbaijan National Academy of Sciences, Azerbaijan
Zeyar Aung, Khalifa University of Science and Technology, UAE
Dr. Shuxiang Xu, University of Tasmania, Australia
Min-Shiang Hwang, Asia University, Taiwan
Thian Song Ong, Multimedia University Malaysia, Malaysia
Dennis Pfisterer, University of Luebeck, Germany
Lalit Prakash Saxena, Combo Consultancy, Obra
Wing Kwong, Hofstra University, USA
Ruchi Tuli, Jubail University college, Jubail, Kingdom of Saudi Arabia
Kai Liu, Chongqing University, China
Petr Sosnin, Ulyanovsk State Technical University, Russia
Vishwas Rudramurthy, Visvesvaraya Technological University, India
Alexandar Djordjevich, City University of Hong Kong, Hong Kong
Mohd Ashraf Ahmad, University Malaysia Pahang, Malaysia
Terje Jensen, Telenor
Shitala Prasad, NTU Singapore, Singapore
Harikumar Rajaguru, Bannari Amman Institute of Technology, Sathyamangalam, India
Vasco N. G. J. Soares, Instituto de Telecomunicações/Instituto Politécnico de Castelo Branco, Portugal
Abdulrahman ALHARBY, Imam Abdulrahman Bin Faisal University, Saudi Arabia
Qiao Xiang, Yale University, USA
Tor-Morten Grønli, Oslo School of Arts, Norway
Hyoungshick Kim, Sungkyunkwan University, Republic of Korea
Robert S Laramee, Swansea University, UK
José Manuel Matos Ribeiro da Fonseca, NOVA University of Lisbon, UNINOVA, Portugal
Shamim H. Ripon, East West University, Bangladesh

Mauro Gaggero, National Research Council of Italy, Italy
Oscar Mortagua Pereira, University of Aveiro, Portugal
Nikos Petrellis, TEI of Thessaly, Larissa, Greece
Suksan Prombanpong, King Mongkut's University of Technology Thonburi, Thailand
Jitender Grover, IIIT-Hyderabad, India
Ahmed Abdelwahab, Qassim University, KSA, Saudi Arabia
Kittisak Jermsittiparsert, Political Science Association of Kasetsart University, Thailand
Seung Yeob Nam, Yeungnam University, Republic of Korea
Marco Listanti, Electronics and Telecommunications, University of Roma Sapienza, Italy
Pascal LORENZ, University of Haute Alsace, France
Reza Malekian, University of Pretoria, South Africa
Yanling Wei, National University of Singapore, Singapore
Ahm Shamsuzzoha, Sultan Qaboos University, Oman
Hyunsung Kim, Kyungil University, Republic of Korea
Filippo Gaudenzi, Università degli Studi di Milano, Italy
Swaroop Joshi, The Ohio State University, USA
Johann M. Marquez-Barja, University of Antwerpen—imec, Belgium
Shuai Zhao, Big Switch, USA
Mohd Faizal Abdollah, Universiti Teknikal Malaysia Melaka, Malaysia
Sallam Osman Fageeri, Al-zaiem Al-azhari University, Khartoum, Sudan
Ng Hui Fuang, Universiti Tunku Abdul Rahman, Malaysia
Alberto Núñez Covarrubias, Universidad Complutense de Madrid, Spain
Wun-She Yap, Universiti Tunku Abdul Rahman
Mehmet Celenk, Ohio University, USA
Salim Ouchtati, Université du 20 Aout 1955—Skikda, Algérie
Cimato Stelvio, Università degli studi di MIlano, Italy
Ximing Fu, Tsinghua University, China

Contents

Part I Ubiquitous Computing

Part II Networks and Information Systems

Part IV Middleware and Operating Systems

Part V Security and Privacy

Part VI Data Mining and Artificial Intelligence

Part VII Software Engineering

Part VIII Web Technology

Part IX Internet of Things

Part I
Ubiquitous Computing

A Study of Wireless Communication Technologies for Vehicular Communication

Afizan Azman, Sumendra Yogarayan, Samuel Leong Wei Jian, Siti Fatimah Abdul Razak, Kirbana Jai Raman, Mohd Fikri Azli Abdullah, Siti Zainab Ibrahim, Anang Hudaya Muhamad Amin and Kalaiarasi Sonai Muthu

Abstract In recent years, vehicle communication is an advanced technology that has attain attention in both industries and academician all over the world. The initiation on vehicular communication is to improve road safety, efficiency and comfort. This paper studies the availability of the wireless communication technologies for vehicular communication and the possible implementation of the suitable wireless communication for vehicle communication in the context of Malaysia.

Keywords Vehicles · Wireless communication · Standards · Spectrum allocation · Alternative

A. Azman (✉) · S. Yogarayan · S. L. W. Jian · S. F. A. Razak · K. J. Raman
M. F. A. Abdullah · S. Z. Ibrahim · A. H. M. Amin · K. S. Muthu
Faculty of Information Science and Information (FIST),
Multimedia University (MMU), Melaka, Malaysia
e-mail: afizan.azman@mmu.edu.my

S. Yogarayan
e-mail: mastersumen@gmail.com

S. L. W. Jian
e-mail: slwj_5319@hotmail.com

S. F. A. Razak
e-mail: fatimah.razak@mmu.edu.my

K. J. Raman
e-mail: jpk_kirbz@hotmail.com

M. F. A. Abdullah
e-mail: mfikriazli.abdullah@mmu.edu.my

S. Z. Ibrahim
e-mail: sitizainab.ibrahim@mmu.edu.my

A. H. M. Amin
e-mail: anang.amin@mmu.edu.my

K. S. Muthu
e-mail: kalaiarasi@mmu.edu.my

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_1

1 Introduction

Vehicle communication have recently drawn great attention, because they have the potential to improve convenience and safety of vehicles. It is possible for vehicles to communicate with each other and not by drivers rolling down their windows to shout across to each other. Vehicle communications can improve the safety, efficiency and productivity of a wide range of commercial vehicles. Using wireless communication technologies, vehicles can communicate via in-vehicle devices that continuously share safety, mobility and environmental information. These connected vehicles can alert drivers to potential hazards, take over vehicle operations when needed, or even manage networks of autonomous vehicles in future.

In Malaysia, two primary technical challenges exist for modeling a vehicular communication. Firstly, the standard of wireless technologies communication is not fixed and standardize for vehicular environment. As such, the limitation of choosing the suitable vehicular communication is high. Secondly, a portion of Malaysian drivers are not aware of the vehicle communication advanced technology. These brings down the interest of research of such a system into vehicles. This research contributes to the development of an integrated platform for vehicle communication. Thus, the technical approaches that are involved are investigated in detail.

2 Literature Review

2.1 Wireless Communication Technologies

In present, there are numerous wireless communication technologies, which could imposed for the vehicular communication. In such, this initiation could bring safety mobility for all. There are a few available wireless technologies, which can be connected for vehicular use, for example, Bluetooth is a wireless technology for exchanging data over short distances between different devices. The chip can be plugged into computers, digital cameras and mobile phones. Bluetooth wireless technology is playing a noteworthy role in car makers [1]. With more cars being manufactured, Bluetooth is installed for hands free calls experience and provide connectivity for music or video files to be transmitted and played in vehicle infotainment system. The drawback is that Bluetooth is not meant for vehicle communication due to its structure and high interference.

Ultra-Wide Band (UWB) is a wireless technology that utilize low power consumption to achieve high bandwidth connections in a communication mechanism. In specific, UWB is meant to transmit huge data over a short distance without using excessive power [2]. UWB could be used in vehicles to exclude the wiring between equipment's and produce high speed of data transmission for information inside and outside of the vehicle. However, there is a concern that UWB implementation in vehicles could end uninvolved vehicles to receive false information.

ZigBee wireless technology is specially designed for sensors and control devices that employ low cost connectivity and widely used for several applications. ZigBee seems to be the key protocol for sensor network applications because of the long battery life, low-cost for installation, eases maintenance and small footprint [3]. Nevertheless, the research of ZigBee technology is quite slow to turn up on the market scale and the coverage are limited with low transmission rate.

Wireless Fidelity (Wi-Fi) is an alternative to wired technology, which is commonly used for connecting devices in wireless mode. Wi-Fi is a generic term that refers to IEEE 802.11 standard for Wireless Local Area Networks (WLANs). Wi-Fi allows to connect any devices to the internet and to each other. Wi-Fi uses radio technologies to transmit and receive data at high speed [4]. Vehicle communication indeed has initiated from the Wi-Fi standard, which includes of IEEE 802.11a/b/g/n. However, in some research has defined the implementation of Wi-Fi is still unstable due to the factor poor efficiency from its physical (PHY) and medium access control (MAC).

On the other hand, Dedicated Short Range Communications (DSRC) is an open-source protocol for wireless communication technology, similar in some respects to Wi-Fi. DSRC is used to enable vehicles to communicate with each other and other road users directly. However, the functionality works well with the standard of licensed spectrum 5.850–5.925 GHz band range which in Malaysia is has not allocated. A combination of DSRC standard IEEE 802.11p and IEE 1609 protocol is set to be denoted as the Wireless Access in Vehicular Environment (WAVE). The standard is a modified version of the IEEE 802.11a and IEEE 1609. WAVE works with low latency, which is essential for vehicular system [5]. However, the implementation of road side units (RSU) is necessary for the functionality of WAVE to completely work as set of providing the safety features. Thus, the cost of implementation will be high.

Cellular network is an underlying technology for mobile phones, personal communication systems and as well wireless technology. The generation of the Global System for Mobile Communication (GSM) is evolve from 2nd Generation (2G), 3rd Generation (3G) and currently at the state of 4th Generation (4G). In today's market conditions car manufacturers evaluate cellular technology as an alternative to provide connectivity in vehicles [6]. However, in terms of vehicle communication, it is not recommended as it uses base station rather than nodes to transmit data. Besides, the communication could expect higher latency compared to the rest of the wireless communication technologies.

In the context of automotive for vehicle communications, Wi-Fi is seems to be a promising technology today due to broad utilization in home, office and public networks and also due to its accessibility. Aside from, Wi-Fi is often used in research projects for vehicle communications. Applications such as drive assistance system reducing collision on road, efficient and improvised traffic control system, communication and information services applications for comfort and convenient to passengers and drivers. In Table 1, several research projects working in this area are shown.

Table 1 Research projects on vehicle

Author	Title	Year	Mechanism
Su, K. C., Wu, H. M., Chang, W. L., and Chou, Y. H	Vehicle-to-vehicle communication system through wi-fi network using android smartphone	2012	Android smartphones with Wi-Fi direct
Jansons, J., Petersons, E., and Bogdanovs, N	WiFi for vehicular communication systems	2013	Wireless base station goodput evaluation
Viittala, H., Soderi, S., Saloranta, J., Hamalainen, M., and Iinatti, J	An experimental evaluation of WiFi-based vehicle-to-vehicle (V2V) communication in a tunnel	2014	Evaluation on Wi-Fi radio in a real tunnel environment
Tornell, S. M., Patra, S., Calafate, C. T., Cano, J. C., and Manzoni, P	GRCBox: extending smartphone connectivity in vehicular networks	2015	Raspberry Pi with Wi-Fi module
Murugesh, R., Ramanadhan, U., Vasudevan, N., Devassy, A., Krishnaswamy, D., and Ramachandran, A	Smartphone based driver assistance system for coordinated lane change	2015	Smartphone extension of Wi-Fi direct
Na, W., Dao, N. N., and Cho, S	Mitigating WiFi interference to improve throughput for in-vehicle infotainment networks	2016	Effective solutions of Wi-Fi interference
Bhawiyuga, A., Sabriansyah, R. A., Yahya, W., and Putra, R. E	A Wi-Fi based electronic road sign for enhancing the awareness of vehicle driver	2016	Raspberry Pi with Wi-Fi module
Vochin, M., and Hayder, A. L	Mobile communication application for V2V systems	2017	Wi-Fi direct

2.2 *Country Standardization*

In late 90s, the U.S Federal Communication Commission in USA has allocated 75 MHz of dedicated short range communication spectrum at 5.9 GHz in range of 5.850–5.925 GHz exclusively for vehicular communication. The purpose of this allocation is to establish safety system that provides safe mobility and efficient traffic. Beside USA, in early 2000s, Japan has also initiated for dedicated short range communication spectrum to allocate 80 MHz at 5.8 GHz in range of 5.770–5.850 GHz [7].

On the other hand, Europe had obstacle to impose vehicular communication standards for safety. This is due to the lacking of dedicated short range communication frequency spectrum. Europe faced a complex and time consuming process in regards to the frequency allocation compared to USA and Japan. In consideration of all the Europe authorities are involved, the steps took a few years for the frequency regulation and deployment. After few years, in 2008, the European Commission has assign

Table 2 Countries standards of spectrum allocation for vehicular communication

Features	USA	Japan	Europe
Radio band (MHz)	75	80	20
Communication range (m)	30	15–20	1000
Communication system	Active and passive	Passive	Passive
Radio frequency (GHz)	5.9	5.8	5.9
Frequency range (GHz)	5.850–5.925	5.770–5.850	5.875–5.905

the spectrum to be attain for vehicular communication. The frequency spectrum is set into range of 5.875–5.905 GHz for road safety at allocation of 20 MHz [8]. In Table 2, the countries standards of spectrum allocation for vehicular communication is tabulated.

In the context of Malaysia, frequency spectrum allocation for vehicular communication is still under a regulatory discussion. The Malaysian Commission has allocated for vehicle activities in range of 5.111–5.268 GHz which the service will not be regarded as a safety service and does not apply to extra-vehicular activities [9]. Thus, the resource of using DSRC in Malaysia is not applicable and realistic. However, the alternative of the DSRC will be Wi-Fi which is still possible to be used in vehicular environment.

3 Proposed Work

Based on the literature review, Wi-Fi shows the most suitable wireless communication protocol to be used compared to others due to its availability and the coverage range. In vehicular environment, throughput rate is an important factor to look into because it is a shared mobility state. Thus, throughput is one of the factor that may result in giving a better result in vehicular communication. Nevertheless, implementing of real-time testing are expensive and network infrastructure is difficult to construct, performance evaluations through simulations are probably a better solution in this work. The main goal of the simulation test is to investigate the possibility of IEEE 802.11n standard for vehicular communication with off the shelf equipment's, available software's and cost effective.

Table 3 Simulation test parameter

Parameter	Value
Distance between node (m)	1–5
Payload size	1472 bytes (1480-8 (header bits))
Simulation time	1–10 s
A-MSDU size	7935 bytes
A-MPDU size	65000 bytes

3.1 Mechanism

Frame aggregation is a feature of the IEEE 802.11n (Wi-Fi) that increases throughput by sending two or more data frames in a single transmission. There are three methods of frame aggregation, MAC Service Data Unit Aggregation (A-MSDU), MAC Protocol Data Unit Aggregation (A-MPDU) and Two Level Aggregation (A-MSDU + A-MPDU). In IEEE 802.11n, the maximum payload per single MAC frame A-MSDU is up to 7935 Bytes and A-MPDU 65,000 Bytes.

3.2 Setup

In order to reach the aim the setup is compared with three aggregation, which is A-MSDU (Scenario A), A-MPDU (Scenario B) and Two-Level Aggregation (Scenario C). In Table 3, the simulation test parameters is tabulated.

In this case, the scenario is a one to one communication where client (Car A) is trying to send a packet to server (Car B). The distance between the nodes for each scenario is configured between 1 and 5 m to test different frame error over distance. For this simulation test, the payload size is set to be 1472 bytes during the transmission and the simulation time is set between 1 and 10 s for each scenario. All the three scenarios will be concurrently running and the throughput is calculated. The throughput adds the total packet transmitted and the payload size over simulation time. Thus, the megabits (Mbits) will be taken as the measure quantity for the throughput.

4 Result

4.1 MAC Service Data Unit Aggregation (A-MSDU)

According to Table 4 tabulated result below, the graph is generated. In Fig. 1, the graph the value increases and decreases as time passed with an average of 2001 packet per second. The throughput value are inconsistent over time. It greatly increases at first 2 s and decrease gradually on 3 s. The pattern value continues the same but with a lower interval gap of increasing and decreasing. On the other hand, it also shows that the throughput are decreasing as the distance are increasing due to higher frame error rate.

4.2 MAC Protocol Data Unit Aggregation (A-MPDU)

According to Table 5 tabulated result below, the graph is generated. In Fig. 2, the graph shows the value increases constantly as time passed with an average of 5053 packet per second. However, the increases value decreases over time. On the other hand, it also shows that the throughput are decreasing as the distance are increasing due to higher frame error rate.

4.3 Two Level Aggregation (A-MSDU + A-MPDU)

According to Table 6 tabulated result below, the graph is generated. In Fig. 3, the graph the value increases as time passed with an average of 5061 packet per second. The throughput value are sharply increases at first 3 s and then it constantly increases later on. However, the increase value decreases over time. On the other hand, it also shows that the throughput are decreasing as the distance are increasing due to higher frame error rate.

Table 4 Simulation test of MAC service data unit aggregation (A-MSDU) result

Time/s	Throughput/Mbits					No of packet
	1 m	2 m	3 m	4 m	5 m	
1	46.5741	23.2870	15.5247	11.6435	9.3148	3955
2	46.6918	23.3459	15.5639	11.6729	9.3383	7930
3	46.8292	23.4146	15.6097	11.7073	9.3658	11930
4	46.8243	23.4121	15.6081	11.7061	9.3648	15905
5	46.8803	23.4401	15.6267	11.7201	9.3760	19905
6	46.8783	23.4391	15.6261	11.7195	9.3757	23885
7	46.9021	23.4510	15.6340	11.7255	9.3804	27880
8	46.9053	23.4526	15.6351	11.7263	9.3811	31865
9	46.9012	23.4506	15.6337	11.7253	9.3802	35845
10	46.9215	23.4607	15.6405	11.7304	9.3843	39845

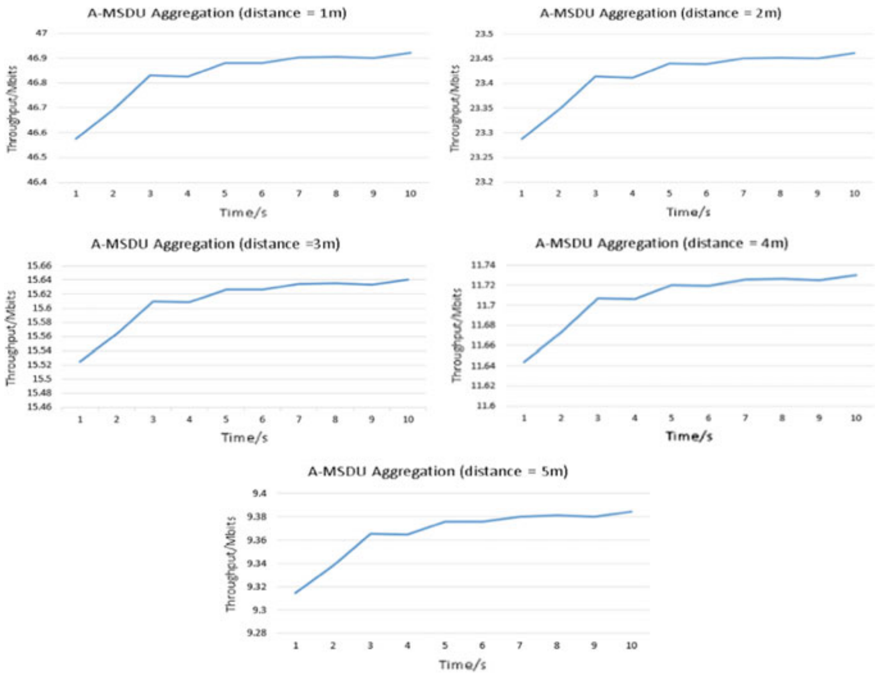


Fig. 1 A-MSDU aggregation graphs

5 Conclusion

The standard of different aggregations scheme in Wi-Fi 802.11n relatively gives an enhancement to the network performance (throughput). Different technique of aggregation are explained based on their parameters. The simulation highlights on different aggregation and how they affect the throughput performance. All the result is explained in term of their capacity packet per transmission. The simulation result of the aggregation shows, Two-Level Aggregation is the highest throughput compared to the others. Although the simulation test is conducted in a close environment, it ensures that there are possibilities of the use of Wi-Fi standard for vehicular communication. Vehicular communication is considered as a technology under development that needs a lot of research and testing.

Acknowledgements This research was carried out within the framework of Connected Car Services Research Group in Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia.

Table 5 Simulation test of MAC service data unit aggregation (A-MPDU) result

Time/s	Throughput/Mbits					No of packet
	1 m	2 m	3 m	4 m	5 m	
1	59.1626	29.5813	19.7209	14.7907	11.8325	5024
2	59.3098	29.6549	19.7699	14.8274	11.8619	10073
3	59.4138	29.7069	19.8046	14.8534	11.8827	15136
4	59.4452	29.7226	19.8151	14.8613	11.8890	20192
5	59.4688	29.7344	19.8229	14.8672	11.8938	25250
6	59.4884	29.7442	19.8295	14.8721	11.8977	30310
7	59.4924	29.7462	19.8308	14.8731	11.8985	35364
8	59.4968	29.7484	19.8323	14.8742	11.8994	40419
9	59.5130	29.7565	19.8376	14.8782	11.9026	45484
10	59.5140	29.7570	19.8380	14.8785	11.9028	50539

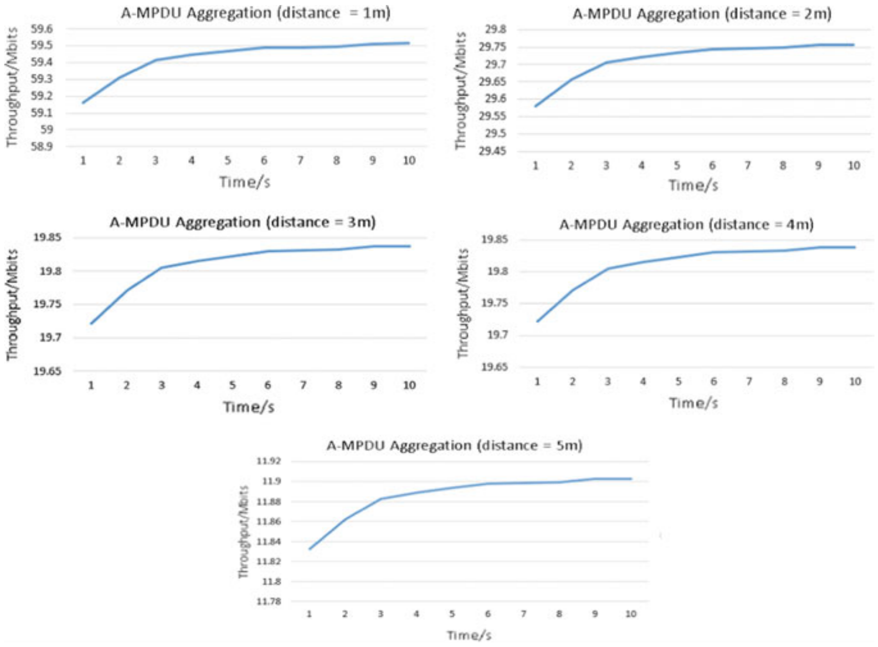


Fig. 2 A-MPDU aggregation graphs

Table 6 Simulation test of two level aggregation (A-MSDU + A-MPDU) result

Time/s	Throughput/Mbits					No of packet
	1 m	2 m	3 m	4 m	5 m	
1	59.7043	29.8522	19.9014	14.9261	11.9408	5070
2	59.7926	29.8963	19.9308	14.9481	11.9582	10155
3	59.8025	29.9012	19.9341	14.9506	11.9605	15235
4	59.8074	29.9037	19.9358	14.9518	11.9614	20315
5	59.8221	29.9110	19.9407	14.9555	11.9644	25400
6	59.8290	29.9145	19.9430	14.9572	11.9658	31050
7	59.8305	29.9152	19.9435	14.9576	11.9661	35565
8	59.8368	29.9184	19.9456	14.9592	11.9673	40650
9	59.8436	29.9218	19.9478	14.9609	11.9687	45720
10	59.8497	29.9248	19.9499	14.9624	11.9699	50615

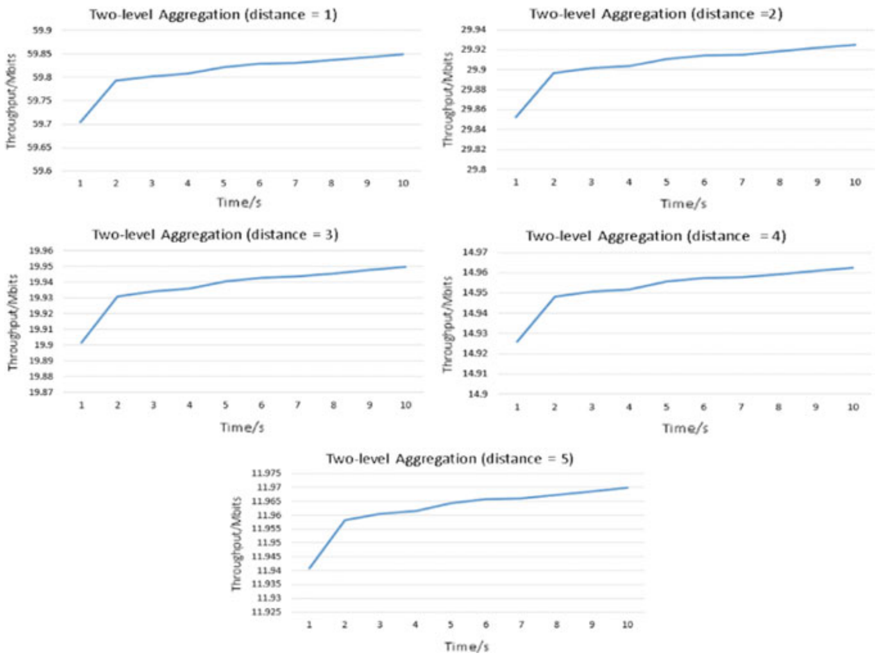


Fig. 3 Two level aggregation (A-MSDU + A-MPDU) graphs

References

1. Nolte T, Hansson H, Bello LL (2005) Wireless automotive communications. In: Euromicro conference on real-time systems, July 2005, vol 6, pp 35–38
2. Zhuang W, Shen XS, Bi Q (2003) Ultra-wideband wireless communications. Wirel Commun Mob Comput 3(6):663–685
3. Dorle SS, Deshpande DM, Keskar AG, Chakole M (2010) Vehicle classification and communication using zigbee protocol. In: 2010 3rd international conference on emerging trends in engineering and technology (ICETET), Nov 2014. IEEE, pp 106–109
4. Banerji S, Chowdhury RS (2013) On IEEE 802.11: wireless LAN technology. arXiv:1307.2661
5. Kenney JB (2011) Dedicated short-range communications (DSRC) standards in the United States. Proc IEEE 99(7):1162–1182
6. Arshad MJ, Farooq A, Shah A (2010) Evolution and development towards 4th generation (4G) mobile communication systems. J Am Sci 6(12):63–68
7. Moustafa H, Senouci SM, Jerbi M (2009) Introduction to vehicular networks. Veh Netw 1
8. Delgrossi L, Zhang T (2009) Dedicated short-range communications. In: Vehicle safety communications: protocols, security, and privacy, pp 44–51
9. Malaysian Communications and Multimedia Commission (2014) Spectrum allocation list—-malaysian communications and multimedia commission (MCMC). https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Spectrum-Plan2014.pdf
10. Su KC, Wu HM, Chang WL, Chou YH (2012) Vehicle-to-vehicle communication system through wi-fi network using android smartphone. In: 2012 international conference on connected vehicles and expo (ICCVE), Dec 2012, pp 191–196
11. Jansons J, Petersons E, Bogdanovs N (2013) WiFi for Vehicular Communication Systems. In: 2013 27th international conference on advanced information networking and applications workshops (WAINA), Mar 2013, pp 425–430
12. Viittala H, Soderi S, Saloranta J, Hamalainen M, Iinatti J (2013) An experimental evaluation of wifi-based vehicle-to-vehicle (V2V) communication in a tunnel. In: 2013 IEEE 77th vehicular technology conference (VTC Spring), June 2013, pp 1–5
13. Tornell SM, Patra S, Calafate CT, Cano JC, Manzoni P (2015) GRCBox: extending smartphone connectivity in vehicular networks. Int J Distrib Sens Netw 11(3):478064
14. Murugesh R, Ramanadhan U, Vasudevan N, Devassy A, Krishnaswamy D, Ramachandran A (2015) Smartphone based driver assistance system for coordinated lane change. In: 2015 International conference on connected vehicles and expo (ICCVE), Oct 2015, pp 385–386
15. Na W, Dao NN, Cho S (2016) Mitigating WiFi interference to improve throughput for in-vehicle infotainment networks. IEEE Wirel Commun 23(1):22–28
16. Bhawiyuga A, Sabriansyah, RA, Yahya W, Putra RE (2017) A Wi-Fi based electronic road sign for enhancing the awareness of vehicle driver. J Phys Conf Ser 801(1):012085
17. Vochin M, Hayder AL. (2017) Mobile communication application for V2V systems. In: 2017 international symposium on signals, circuits and systems (ISSCS), July 2017, pp 1–4

A New Fractional Binary-to-Decimal Number Conversion for Errorless Calculation

Jinhyuck Kim and Hyun Jin Kim

Abstract This paper proposes a new fractional decimal format and binary-to-decimal conversion for highly accurate calculation without error. In many cases, a given fractional constant is the decimal number. After converting the decimal constant to a fractional binary number, any calculation can be performed. Because the given constant is originally the decimal number, the result may contain an error, which degrades the system accuracy. We propose a new format for fractional decimal numbers. The proposed format contains a binary integer and fractional bits, where the fractional decimal number can be provided by the proposed binary bit conversion. In a calculation, instead of using fixed-point numbers, scaled integer numbers are adopted. Then, the result is scaled and converted into the fractional decimal number. Considering practical examples and evaluation results, it is concluded that the proposed method can provide the errorless calculation.

Keywords Binary-to-decimal conversion · Decimal
Fixed-point number · Fractional binary number

1 Introduction

In many embedded systems, a real number is represented in the form of a fractional binary number [1]. In general, there are two methods of expressing a fractional binary number: the floating-point and fixed-point representations. Compared to the fixed-point representation, the floating-point representation has both a larger numbering range and enhanced accuracy. However, it incurs a large hardware overhead. In addition, the computing speed is slow due to the complex hardware structure. Therefore, the fixed-point representation is mainly adopted in small embedded systems, where the hardware calculation is performed in an arithmetic unit using fixed-

J. Kim · H. J. Kim (✉)
School of EEE, Dankook University, Yongin-si, Gyeonggi-do, Republic of Korea
e-mail: hyunjin2.kim@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_2

format binary numbers [2]. When converting a decimal fractional constant into a fixed-format binary number, a conversion error can occur [3].

In this paper, to eliminate this conversion error, a new binary fractional format and binary-to-decimal conversion method are proposed. The new format of the fractional binary number contains integer and fractional related bits, which are used to obtain the fractional decimal number by the proposed conversion method. In the proposed conversion, the integer value is calculated by scaling the fractional number. Considering practical examples and evaluation results, it is concluded that the proposed format and conversion method can provide the errorless calculation with low hardware overhead.

2 Proposed Fraction Decimal Format and Conversion

2.1 Motivation

In general, when converting a decimal fraction into a fixed-point binary representation, a conversion error can occur. For example, to represent the decimal fraction 2.285 with a 10-bit fixed-point binary representation, the following equation can be used:

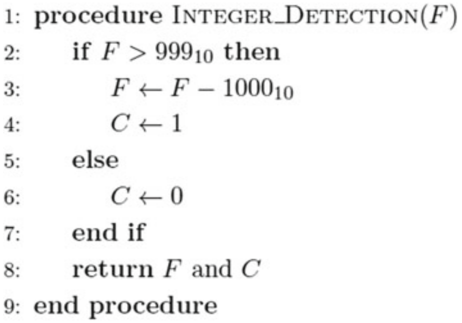
$$2.285_{10} = 2^1 + 2^{-2} + 2^{-5} + 2^{-9} + 2^{-10} + \cdots .$$

Therefore, the value of the fixed-format binary number cannot be the same as 2.285_{10}. The value 2.285 can be represented as 2,285 in a fixed-point data type with a scaling factor of $\frac{1}{1000}$. Instead of using the fixed-format binary number, if the unscaled binary integer of 2,285 is adopted in operations, the conversion error can be avoided. Therefore, based on the motivation of the unscaled binary number without any conversion error, we propose a new format of binary fractional number and its binary-to-decimal conversion to obtain the fractional decimal number.

2.2 New Format for Binary Fractional Number

To show the new format for binary fractional numbers, a positive real number with three decimal digits X with a scaling factor of $\frac{1}{1000}$ is assumed. The number Y is converted into X with 16 bits. In the new format, because $\frac{2^{16}}{1000}$ is 65.535_{10}, the range of a real number X is from 0 to 65.535_{10}.

Due to the three decimal digits, $X[9{:}\ 0]$ is defined as the *fractional related bits*. In the proposed format, if the value stored in the fractional related bits is smaller than 1000_{10}, the value of the fractional related bits after scaling is also less than one. Otherwise, the fractional related bits mean a mixed number of $1.f$, where f



```
procedure INTEGER_DETECTION($F$)
    if $F > 999_{10}$ then
        $F \leftarrow F - 1000_{10}$
        $C \leftarrow 1$
    else
        $C \leftarrow 0$
    end if
    return $F$ and $C$
end procedure
```

Fig. 1 Pseudocode of integer detection

means a fraction. On the other hand, X[15: 10] is defined as the *integer related bits*. According to the value of X, the integer related bits can be a mixed number with integer values and a fraction. The details of the conversion are explained in the following subsection.

2.3 Proposed Binary-to-Decimal Conversion

Both the fractional and integer related bits can be mixed numbers with an integer value and fraction. By using the proposed conversion, the fractional decimal number is obtained. Firstly, the fractional related bits can be a mixed number with non-zero integer value and a fraction. The pseudocode of the procedure INTEGER_DETECTION is shown in Fig. 1. When the value of the fractional related bits is greater than 999_{10}, F is a mixed number with non-zero integer value of one. Therefore, C is set to one and then returned. Otherwise, the returned value of C is zero. In addition, the fraction F is also returned.

Secondly, the integer related bits can also be a mixed number with integer value and a fraction. For the conversion of the integer related bits, the pseudocode of the procedure FRACTION_DETECTION is shown in Fig. 2. Procedure FRACTION DETECTION returns the fraction from the integer related bits when I is a mixed number. In the code, the fraction f is initialized to 0. Then, each bit of the integer related bits is compared with one. If the bit is one, its own fractional number is added to f. The pseudocode in Fig. 2 shows the fraction detection rule for each bit of the integer related bits. For example, $I[5]$ is equal to $X[15]$, which means that $2^{15} = 32768_{10}$ to be scaled with a factor of $\frac{1}{1000}$. Therefore, 768_{10} or 1100000000_2 can be the binary number for the fraction. For $I[4]$ and $I[3]$, the same method as that described above is applied. For $I[2{:}\ 0]$, **switch** statement is used for clarity.

On the other hand, if the integer related bits are greater than 101001_2 or 42_{10}, f is a mixed number with non-zero integer value. In the conversion, therefore, the integer value from FRACTION_DETECTION should be considered.

Fig. 2 Pseudocode of fraction detection

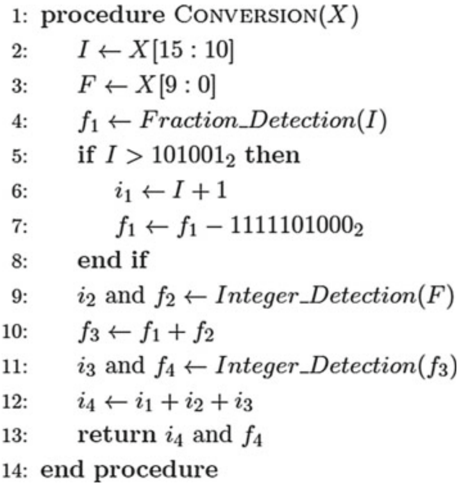
```
procedure FRACTION_DETECTION($I$)
    $f \leftarrow 0$
    if $I[5] = 1$ then $f \leftarrow f + 1100000000_2$
    end if
    if $I[4] = 1$ then $f \leftarrow f + 110000000_2$
    end if
    if $I[3] = 1$ then $f \leftarrow f + 11000000_2$
    end if
    switch I[2:0] do
        case $000_2$ $f \leftarrow f + 0_2$
        case $001_2$ $f \leftarrow f + 11000_2$
        case $010_2$ $f \leftarrow f + 110000_2$
        case $011_2$ $f \leftarrow f + 1001000_2$
        case $100_2$ $f \leftarrow f + 1100000_2$
        case $101_2$ $f \leftarrow f + 1111000_2$
        case $110_2$ $f \leftarrow f + 10010000_2$
        case $111_2$ $f \leftarrow f + 10101000_2$
    return $f$
end procedure
```

Using the procedures INTEGER_DETECTION and FRACTION_DETECTION, the proposed conversion is explained in the procedure CONVERSION in Fig. 3, where i_1, i_2, i_3, i_4 and f_1, f_2, f_3, f_4 mean the temporary variables for the integer value and fraction, respectively. In the conversion, the values of the integer and fraction are obtained for a newly formatted 16-bit binary number X. Firstly, the input bits X are divided into the integer related bits I and fractional related bits F. Then, the fraction f_1 is returned by FRACTION_DETECTION. If the integer related bits are greater than 101001_2, f_1 can be a mixed number with non-zero integer of one and fraction.

Therefore, when I is greater than 101001_2 or 42_{10}, i_1 is set to $I + 1$. In addition, to subtract the integer of one from f_1, 1111101000_2 or 1000_{10} is subtracted. Next, from INTEGER_DETECTION, the values of the integer and fraction for F are obtained. The temporary variable f_3 is obtained by summing f_1 and f_2. Then, by calling INTEGER_DETECTION again, the values of the integer and fraction for f_3 are obtained. Finally, the integer and fraction results i_4 and f_4 are returned, respectively.

2.4 Example of Proposed Conversion

Examples of the proposed conversion are described using the notations in Figs. 1, 2, and 3 as follows: the first example is the conversion of 56.759_{10}. In this case, X is



```
1: procedure CONVERSION($X$)
2:     $I \leftarrow X[15:10]$
3:     $F \leftarrow X[9:0]$
4:     $f_1 \leftarrow Fraction\_Detection(I)$
5:     if $I > 101001_2$ then
6:         $i_1 \leftarrow I + 1$
7:         $f_1 \leftarrow f_1 - 1111101000_2$
8:     end if
9:     $i_2$ and $f_2 \leftarrow Integer\_Detection(F)$
10:    $f_3 \leftarrow f_1 + f_2$
11:    $i_3$ and $f_4 \leftarrow Integer\_Detection(f_3)$
12:    $i_4 \leftarrow i_1 + i_2 + i_3$
13:    return $i_4$ and $f_4$
14: end procedure
```

Fig. 3 Pseudocode of conversion

56.759_{10} or 1101110110110111_2, where I and F are 110111_2 and 0110110111_2, respectively. Firstly, I is input into FRACTION_DETECTION. For the FRACTION_DETECTION, $I[5]$, $I[4]$, and $I[2:0]$ are 1_1, 1_1, and 111_1, respectively. Therefore, in line 4 of Fig. 3, the value of f_1 returned by FRACTION_DETECTION is 10100101000_2, which is calculated as $1100000000_2 + 110000000_2 + 10101000_2$. Because I is greater than 101001_2, i_1 is set to 111000_2. In addition, f_1 can be 101000000_2, which is calculated as $f_1 - 1111101000_2$. Next, F is input into INTEGER_DETECTION. Because F is not greater than 1111101000_2, i_2 is 0. In addition, f_2 is equal to F or 0110110111_2. By adding f_1 and f_2, the fraction related bits f_3 is obtained. After f_3 is input into the second INTEGER_DETECTION, i_3 and f_4 are 0 and 011011011_2, respectively. Finally, i_4 is 111000_2, which is the same as i_2 because i_3 and i_4 are both 0. In addition, f_4 is 1011110111_2. In the newly formatted X, the decimal fractional value 56.759_{10} is represented as $56, 759_{10}$ in the fixed-point data type with a scaling factor of $\frac{1}{1000}$. After performing the proposed conversion, the returned values i_4 and f_4 are 56_{10} and 759_{10}, respectively.

3 Evaluation of Proposed Format and Conversion

3.1 Comparison of Error with Fixed Point Binary Format

To compare with the fixed-point binary format in terms of error, errors were calculated from many fixed-point binary numbers, which contained 10 fraction bits. In this evaluation, the decimal numbers that had the decimal point consisting 3 digits were assumed to be represented with the proposed format and fixed-point binary format, respectively.

Table 1 Comparison of implementation

	Proposed format	Fixed-point binary format
Total combinational functions	114	880
Dedicated logic registers	35	69
$F_{\max}$ (MHz)	68.54	29.84

In the calculation of fixed point numbers in 1000 cases within from 0.000_{10} to 0.999_{10}, the maximum error was 0.000976_{10}. On average, the fixed-point numbers had an error of 0.000486_{10}. Therefore, the fixed-point format incorporated an error after the decimal point within the same range. Because there was no error in the proposed format and binary-to-decimal conversion within the allowed range, it is concluded that the proposed format affords errorless calculation unlike the fixed-point binary format.

3.2 Comparison with Fixed Point Format in Terms of Hardware Overhead and Maximum Speed

For an apples-to-apples comparison, the conversion for the proposed and fixed-point formats were coded and then implemented with Altera's CycloneII EP2C20F484C7 device [4]. The implementations provided BCD (binary-coded decimal) codes after one clock cycle for the proposed formatted and fixed-point numbers, respectively. The hardware overhead and F_{max} are shown in Table 1.

As shown in Table 1, the proposed format conversion provided a high operating frequency with low hardware overhead. The reason for the performance enhancement with lower hardware overhead can be discussed as follows: in the fixed-point format, to get the BCD code for the fractional binary number, a hardware conversion module was required. On the other hand, the proposed conversion adopted the unscaled binary integer format, which was directly used to get the BCD code for the fractional number with low hardware overhead.

4 Conclusion

The proposed fractional decimal format and binary-to-decimal conversion provide for errorless calculation. Compared to the general fixed-point number, a decimal number can be converted into its own binary number without any conversion error. For given operation, instead of using the fixed-format binary number, the unscaled binary integer is adopted. Then, the result is scaled and converted into the fractional decimal number. Therefore, the new format can be applied to any arithmetic operation. Considering the examples mentioned above, the proposed format and conversion can provide for binary-to-decimal conversion without any conversion error. Considering the evaluation results mentioned above, it is concluded that the proposed format can provide BCD codes using both integer and fractional binary numbers with low hardware overhead.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education. (2017R1D1A1B03030348).

References

1. Gordon Robert (1998) A calculated look at fixed-point arithmetic. Embed Syst Program 11(4):72–79
2. Introduction to fixed point number representation. https://inst.eecs.berkeley.edu
3. Tocci R (2010) Digital systems: principles and applications. Pearson Education
4. Cyclone II devices homepage. http://www.altera.com

Development of Distraction Limit Estimation Index Using Posture Change Monitoring System

Yun-Hong Noh, Ji-Yun Seo and Do-Un Jeong

Abstract Modern people spend a lot of time in sitting on the chair and sofa, because of watching TV, studying, work, driving. However, when sitting for a with inappropriate posture for a long time, musculoskeletal disorder is induced, frequent movement increases distraction, giving unreasonable work and studying at low concentration. Therefore, it is very important to maintain the proper posture, and a system that guides the user to a correct posture when sitting in a wrong posture is required. In this research, a cushion type monitoring system based on sitting information is implemented and it is possible to monitor distraction and health management by posture detection. The implemented system uses 8 pressure sensors and it can judge the normal posture and 8 kinds of wrong posture and feed back to the correct posture to the user. As a result of wrong postures and a correct posture discrimination results, it showed discrimination performance of 98%.

Keywords Musculoskeletal disorder · Posture correction · Monitoring system

1 Introduction

Modern society is suffering from various kinds of spinal diseases due to long sitting life, and most of the causes of the disease are in the wrong posture habits. When you sit for a long time in the wrong position, the vertebrae and the pelvis give a lot of strain. Therefore, it is very important to detect the wrong posture early and to guide it to the correct posture. In previous research, the method of attaching the

Y.-H. Noh
Busan Digital University, 57 Jurye-ro, Sasang-gu, Busan 47011, Korea
e-mail: yhnoh@bdu.ac.kr

J.-Y. Seo · D.-U. Jeong (✉)
Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 47011, Korea
e-mail: dujeong@dongseo.ac.kr

J.-Y. Seo
e-mail: 92sjy02@naver.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_3

sensor to the form of chair, bed, and toilet was studied for the posture correction in daily life [1, 2]. In particular, chairs are one of the most developed apparatuses for modern people who spend the most seating time. However, chair-type systems with posture correction are difficult to use in everyday life due to portability and cost. In this research, we implemented smart cushion without restraint, which is easy to apply to existing chair and has health management function of posture correction and distraction monitoring. The implemented system continuously monitors the weight distribution information and can feedback the wrong posture and habit information to the user to judge the habit of inducing correct posture and reducing the concentration on learning and work. For this purpose, a total of eight pressure sensors are arranged and a system control section based on Arduino is constructed.

2 System Configuration

In this research, we implemented a measurement section consisting of eight pressure sensors in the sitting area to amplify the signals according to the weight distribution of the seated person. Measuring sensor is FSR-406 pressure sensor of Interlink Company was used for measurement. The implemented system measures the seating information of the user, and then calculates accuracy and error based on the standardized data to determine the seating posture of the user. Thereafter, when the normal posture is maintained, the user maintains the posture through continuous monitoring. In the case of abnormal posture, the system is guided to the normal posture through application Alarm. We implemented a system control section using Arduino and used Bluetooth transmission system for transmitting to a smart phone. The implemented system configuration is shown in Fig. 1.

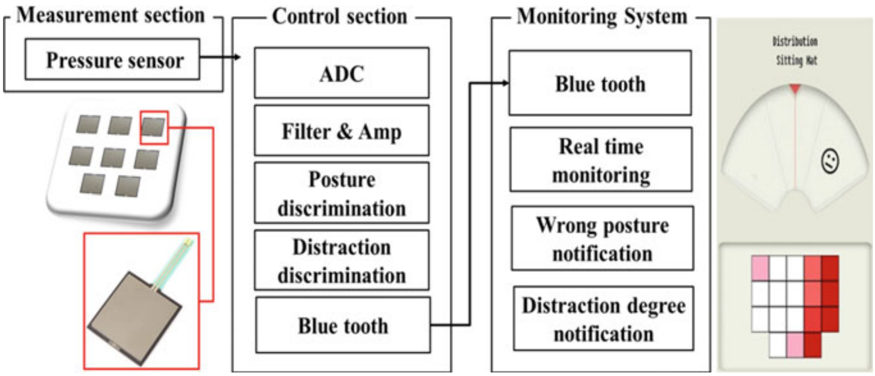


Fig. 1 System configuration

3 Detection of Posture and Distraction

3.1 Posture Detection

In this research we implemented triangle center algorithm of weight information base for posture and distraction detect shown in Eq. 1, the center of gravity of the signal output from the eight pressure sensor was calculated (left = a, right = b, rear = c).

Using the triangular center algorithm, it is possible to confirm the intensity and frequency of movement reflecting the detection of the center of gravity movement, the information of posture change and the distance value. In order to measure the intensity and frequency size of movement, the distance between the current center on the coordinate and the previous center was calculated, and the value of the distance was reflected in the intensity and frequency of the posture change.

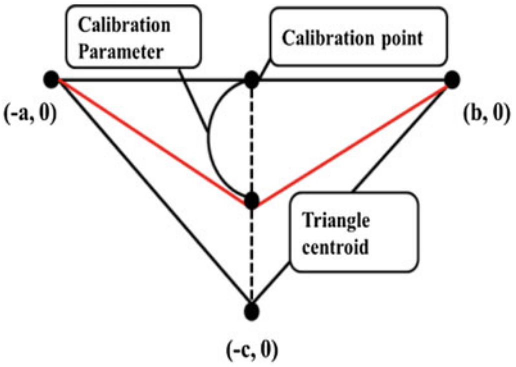
$$TC_1(x, y) = \left(\frac{-a_1 + b_1 + 0}{3}, \frac{0 + 0 - c_1}{3}\right) - \left(X_{cp} Y_{cp}\right) \tag{1}$$

Equation 2 and Fig. 2, shows the distance calculation in the coordinate. TC is used at the setting position of the central coordinate reflected at the time of seating at the beginning, and initializes the coordinates generated by subsequent posture change. The center point (TC_1) is represented by a reference point (CP) for checking the degree of change of the center point (TC_2) continuously detected. For posture discrimination, apply the current center point to $\left(\mathrm{x}_1, \mathrm{y}_1\right)$ apply the previous coordinates to $\left(\mathrm{x}_2, \mathrm{y}_2\right)$ and set the straight line value of the shortest distance between both centers to calculated and applied.

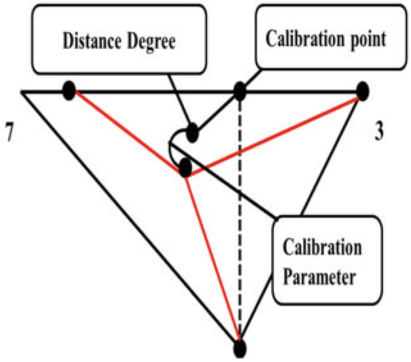
$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2}$$

3.2 Distraction Detection

Existing distraction detection has been researched extensively using EEG. However, EEG has many problems ordinary people to use in everyday life. In order to solve this problem, there are some studies that measure the degree of distraction based on the movement of the user [3]. It research's determined that the user is in a concentrated state as the movement value is smaller. That is, even if you are not using EEG, you can measure the degree of distraction that is the inverse relation of concentration. In this research, distraction degree is calculated by reflecting the distance and the time between two center points caused by the center of gravity movement determined by the implemented algorithm. In Eq. 3, the DLE index (Distraction-Limit-Estimation index) is expressed, and the distance is reflected in the distance between the center



(a) Movement value generation and Center point movement



(b) First triangle center discrimination and Standard parameters

Fig. 2 Distance calculation in the coordinate

points before and after the movement, and the time domain of the detected peak is reflected in time. The reflection of the time domain computes n areas, which is the cumulative number of frequency and intensity, and utilizes as a DLE Index.

$$DLE\ Index_n = \sum_{k=0}^{n} (Distraction \times Time)_k \tag{3}$$

4 Experiments and Results

4.1 Posture Measurement Evaluation

For the performance evaluation of the implemented system, nine sitting posture of the experimenter were measured 50 times in total. The nine measured postures are (a)

Table 1 Posture determination performance

		Take posture								
		a	b	c	d	e	f	g	h	i
Detection	a	50								
	b		50							
	c			50						
	d				50					
	e					47			1	
	f						47			
	g							48		
	h					1	1		49	
	i									49

Left trembling, (b) right trembling, (c) left tilt, (d) right tilt, (e) left twisting, (f) right twisting, (g) front tilt, (h) back tilt, (g) normal posture. The performance evaluation results of the implemented system are shown in Table 1.

As a result, the detection performance of the normal posture, left right leg trembling posture and tilt posture, was 100%, left and right leg twisting posture 94%, front and back tilt posture 97%. Therefore, it was confirmed that the total detection success rate is 98%, and the posture can be judged through the sitting information.

4.2 Distraction Detection Performance Evaluation

In order to evaluate the performance of the implemented system, 10 healthy college students were evaluated for distraction according to the amount of posture change when they watched audio-visual data for 10 min. In order to measure the degree of distraction, a DLE index based on the intensity and frequency of the posture change was implemented. The 10 experimental subjects showed the average intensity of the average of 18 posture changes during 10 min of watched audio-visual. When the experimenter watched the audiovisual data, the posture change was most affected by the habit. As a result of the experiment, the value of DEL index increased with increasing intensity and frequency, and the value of standard deviation decreased with increasing DLE index. This means that the greater the intensity and frequency of change of posture, the lower the concentration. The results of the DLE index evaluation test are shown in Table 2.

Table 2 Posture and distraction data captured by the proposed system for 10 subjects

Subjects	Proposed system summary statistics			
	Detection using camera		DLE index in posture change	Standard deviation
	Intensity	Posture changes		
1	3	18	21.6	0.71
2	4	24	38.4	0.41
3	2	16	12.8	0.85
4	1	15	6.2	0.89
5	3	20	24.1	0.73
6	2	21	16.8	0.81
7	4	23	36.8	0.45
8	5	24	48.3	0.13
9	2	9	7.2	0.88
10	3	18	25.6	0.69

5 Conclusion

In this research, we implemented a cushion type monitoring system based on seating posture information to improve concentration. The implemented system continuously monitors the weight distribution information and can feedback the wrong posture and habit information to the user to judge the habit of inducing correct posture and reducing the concentration on learning and work. Using the triangular center algorithm, it is possible to confirm the intensity and frequency of movement reflecting the detection of the center of gravity movement, the information of posture change and the distance value. Distraction degree is calculated by reflecting the distance and the time between two center points caused by the center of gravity movement determined by the implemented algorithm. For the performance evaluation of the implemented system, nine sitting posture of the experimenter were measured 50 times in total. It was confirmed that the total posture detection success rate is 98%, and result of analyzing the degree of distraction due to posture change, it was confirmed that the higher the degree of distraction, the decreased the standard deviation. This means that the greater the intensity and frequency of change of posture, the lower the concentration. In future research, we will investigate the correlation with EEG for more accurate posture detection and concentration analysis.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01061131, No. 2016R1D1A1B03934866).

References

1. Ostadabbas S et al (2014) In-bed posture classification and limb identification. In: 2014 IEEE biomedical circuits and systems conference (BioCAS). IEEE
2. Lee H et al (2017) ADHD assessment and testing system design based on virtual reality. In: 2017 2nd international conference on information technology (INCIT). IEEE
3. Hansen JHL et al (2017) Driver modeling for detection and assessment of driver distraction: examples from the UTDrive test bed. IEEE Signal Process Mag 34:4

Implementation of Rehabilitation Assistant System Based on Movement and Muscle Activity Information

Ji-Yun Seo, Yun-Hong Noh and Do-Un Jeong

Abstract Existing rehabilitation treatment is experience base of experts, to do a lot of treatment and training. However, in this research, we implemented a rehabilitation support system based on movement and muscle activity data that can support efficient rehabilitation based on more objective data. Implemented system utilizes EMG, acceleration sensor and gyro sensor, it becomes a measurement, so it is possible to accumulate more objective data and plan a treatment when doing rehabilitation treatment. In order to evaluate the performance of the implemented system, we measured EMG data and movement data were measured assuming femoral muscle related rehabilitation exercise situations. As a result of the experiment, four situations classifications were possible and comparative evaluation with commercial systems also confirmed very similar results.

Keywords Rehabilitation treatment · EMG · Acceleration sensor

1 Introduction

In modern society that develops rapidly, the number of patients due to traffic accidents, occupational accidents, etc. is increasing every year, and the increase in the elderly population, which is a social problem, and the chronic diseases related to adult diseases tend to increase [1, 2]. In addition, secondary disease caused by unintended accidents and genetic problems are also increasing. Such patients need rehabilitation treatment to return to their daily lives, but rehabilitation treatment is often treated to

J.-Y. Seo · D.-U. Jeong (✉)
Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 47011, Korea
e-mail: dujeong@dongseo.ac.kr

J.-Y. Seo
e-mail: 92sjy02@naver.com

Y.-H. Noh
Busan Digital University, 57 Jurye-ro, Sasang-gu, Busan 47011, Korea
e-mail: yhnoh@bdu.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_4

therapist's experience and intuition, and systematic treatment is not done [3]. In this research, we implemented EMG based muscle momentum estimation algorithm to make systematic and objective planning and treatment more than conventional therapy. The implemented system can simultaneously monitor the activity potential of the muscles and the movement information of the user to objectively present the effect of the rehabilitation exercise. Various movement parameters are extracted via an algorithm and the extent of movement accompanying actual muscle use is measured to judge momentum and effect. In addition, by implementing a smartphone-based monitoring system section for real-time monitoring, it can help efficient rehabilitation treatment and manage long-term user's rehabilitation treatment.

2 System Configuration

Previously, the rehabilitation therapy using EMG was used for the purpose of checking the condition of the patient without using the EMG data of the patient for treatment or rehabilitation exercise. In other words, it was used to prescribe customized rehabilitation exercise for each patient. However, the implemented system is a medical assistive device that enables a more efficient rehabilitation exercise by assisting the rehabilitation treatment, not the measurement and the prescription for the patient's condition. The implemented system consists of a sensor based measurement section for measuring body movement information and a smartphone based monitoring system for processing the measured data to extract muscle momentum and motion parameters has been done. The system configuration is shown in Fig. 1.

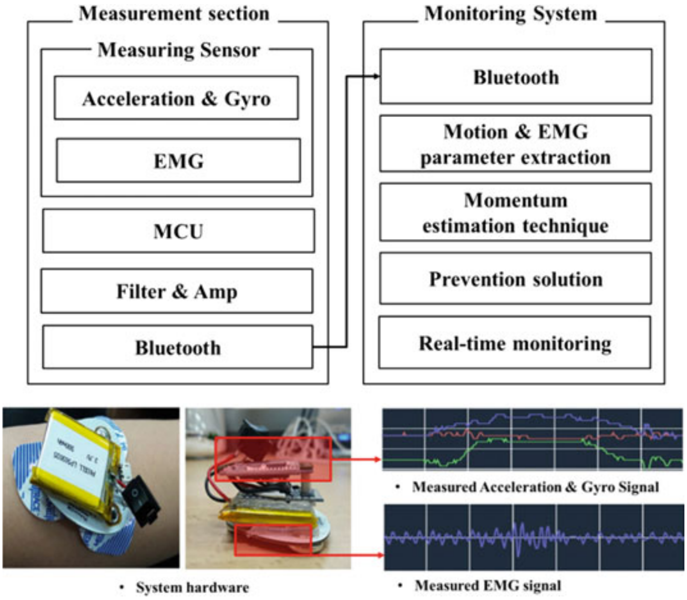
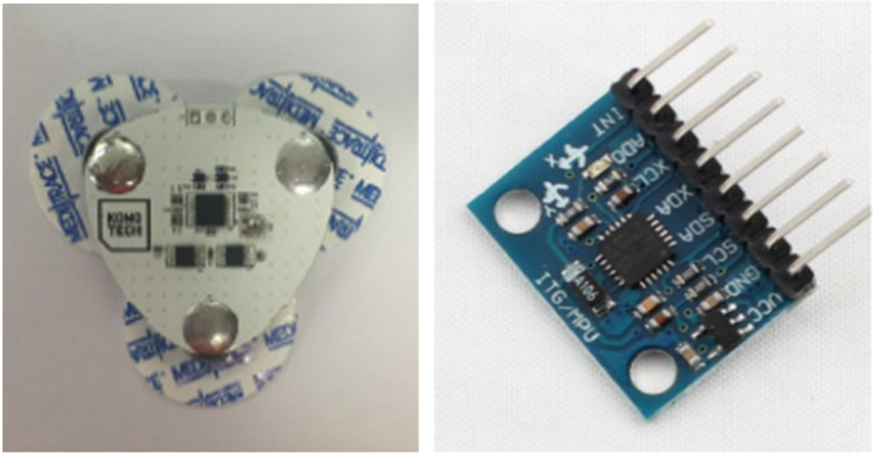


Fig. 1 System configuration

3 Measurement Section

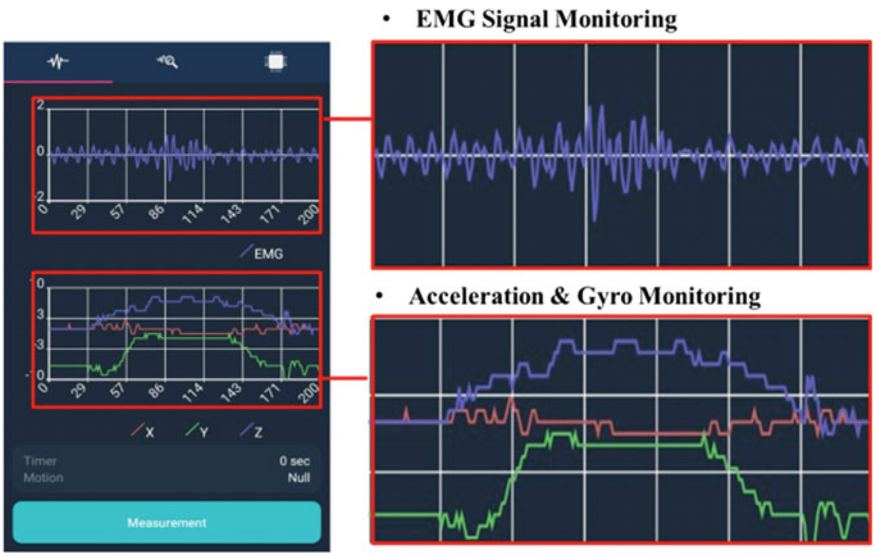
The hardware system implemented in this research consists of a measurement section that measures myoelectric potential, a control section to analyze and process the measured signal, a transmission that transmits the processed signal via the Bluetooth module. We used acceleration sensor and gyro sensor to judgment the user's movement. 3-axis acceleration sensor and gyro sensor simplified the design of measurement system using integrated MPU 6050. By using only the acceleration sensor and the gyro sensor, it is possible to know whether the user is performing a rehabilitation exercise. However, in this case, we cannot know how much muscle activities are used for the rehabilitation exercise, so we used the EMG sensor together to measurement the user's movement and muscle activity information simultaneously. For measurement of EMG, we are using a small wearable EMG module of Kong-Tech Company, which is designed to be used for attaching to the arm, but attached to various body parts there are possible features. We used ATmega 328 based Arduino Pro Mini to process measured sensor information. Perform the function of collecting and analyzing data using the EMG module, acceleration sensor, gyro sensor and transferring the result to the smartphone via Bluetooth communication. Particular, in this research, considering the case of attaching a large number of sensors, we applied HC-06 modules which make it easy to interface with Arduino by applying Bluetooth communication technology for multiple sensors and smartphone interface did (Fig. 2).



(a) EMG sensor module

(b) MPU 6050 sensor

Fig. 2 Sensor used in hardware measurement section



Timer 0 sec
Motion Null

Fig. 3 Implemented monitoring system application

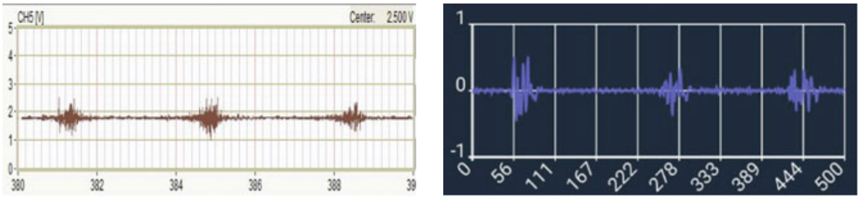
4 Monitoring System

We implemented an application to save and analyze the transmitted EMG information, acceleration sensor and gyro sensor values. The implemented smartphone-based monitoring system has an algorithm that can distinguish the momentum and posture of muscles. In order to provide posture and exercise information that can maximize the treatment effect during rehabilitation exercise, in the current research, we focused on discrimination function of exercise posture (Fig. 3).

5 Experiments and Results

5.1 EMG Comparative Evaluation Experiment with Commercial System

For more objective experiments, we compared physiolab company commercial EMG measurement system with the implemented EMG measurement system. The experiment took the same movement on both systems and analyzed the correlation of the measured EMG data. As a result of EMG measurement, we confirmed that the results of the commercial system and the implemented system are very similar (Fig. 4).



(a) EMG measurement of commercialization system

(b) EMG measurement of implemented system

Fig. 4 EMG signal comparison between systems

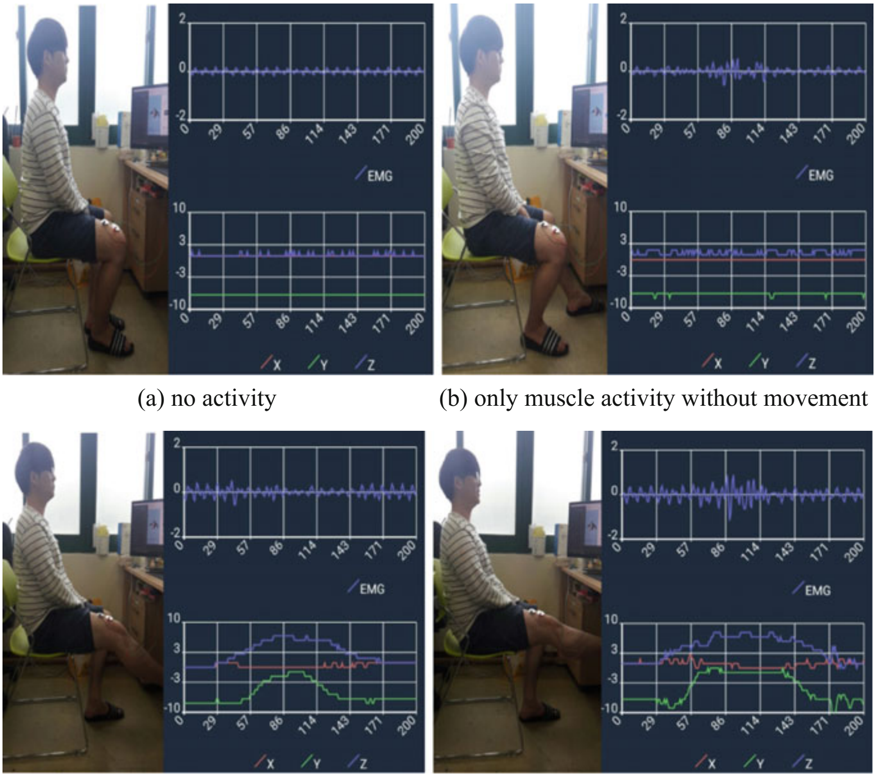
5.2 Performance Evaluation of Implemented Systems

There are many problems to determine whether rehabilitation exercise is effective by using EMG data only. Therefore, for effective rehabilitation, movement information and muscle activity information should be measured simultaneously. In this research, to obtain an effective rehabilitation exercise, we measured movement information and muscle activity information at the same time to determine correct exercise posture. To evaluate the performance of the implemented system, EMG data and movement data were measured assuming femoral muscle related rehabilitation exercise situations (Fig. 5).

In the case of experiment (a), there is no change in the graph, since no action is taken. In the case of (b), the EMG signal is measured because of muscle activity, but there is no change in acceleration and gyro graph because there is no movement. (c), contrary to (b), there was movement but no muscle activity. Therefore, the EMG graph does not change, and the acceleration and the gyro graph show that the Y axis and the Z axis rise and then decrease. (d) is an experimental posture and a measured data graph when the user sits in a right posture and lifts the leg up and down. In the experiment (d), the EMG signal was observed in the graph, and in the case of the acceleration and the gyro graph, the value of the Y axis and the Z axis increased and then decreased. Therefore, if the graph appears as (d), the user can judge the correct rehabilitation exercise posture.

6 Conclusion

In this research, we developed a system to judge whether or not we are exercising in a correct posture during rehabilitation exercise. By using only the acceleration sensor and the gyro sensor, it is possible to know whether the user is performing a rehabilitation exercise. However, in this case, we cannot know how much muscle activities are used for the rehabilitation exercise, so we used the EMG sensor together to measurement the user's movement and muscle activity information simultaneously. To



(a) no activity

(b) only muscle activity without movement

(c) only movement without muscle activity

(d) Correct rehabilitation activity

Fig. 5 4 measurement situations for experiment

evaluate the performance of the implemented system, EMG data and movement data were measured assuming femoral muscle related rehabilitation exercise situations. As a result of the experiment, four posture classifications were possible, there were differences in electromyogram values depending on the physique, but in all other postures it was all possible to make a judgment. In future research, the system will be applied to actual rehabilitation treatment to carry out continuous research to more objectively evaluate the effect of rehabilitation treatment.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01061131, No. 2016R1D1A1B03934866).

References

1. Ellington A et al (2015) Behavioral intention to use a virtual instrumental activities of daily living system among people with stroke. Am J Occup Therapy 69(3):6903290030p1–6903290030p8
2. Liu L et al (2017) Development of an EMG-ACC-based upper limb rehabilitation training system. IEEE Trans Neural Syst Rehabil Eng 25(3):244–253
3. Cao W-J et al (2017) Study on a novel wearable exoskeleton hand function training system based on EMG triggering. In: Wearable Sensors and Robots. Springer, Singapore, pp 135–143

Part II
Networks and Information Systems

NUPT ST-Data Miner: An Spatio-Temporal Data Analysis and Visualization System

Zhiqiang Zou, Junjie Xiong, Xu He and Haihong Dai

Abstract Given the increasing popularity and availability of location tracking devices, large quantities of Spatio-Temporal data (ST-data) are available from many different sources. For the ST-data, reflecting the mobile characteristic of the world, it is essential to build a functional system to perform quickly interactive analysis. In this paper, we present an analysis and visualization system, NUPT ST-data Miner, which facilitates users to visualize and analyze ST-data. It (1) provides a flexible and extensible framework based on cloud computing platform, (2) is able to quickly retrieve specified ST-data, (3) integrated multiple functions for the ST-data. To demonstrate its efficiency, we validate our model and system on a real data set of Microsoft Research Asia. The results from extensive experiments demonstrate that NUPT ST-data Miner is an effective system for visually analyzing spatio-temporal data.

Keywords Spatio-Temporal analysis · Visualization · Big data · Cloud computing · GIS

Z. Zou · J. Xiong (✉) · X. He · H. Dai
College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210023, People's Republic of China
e-mail: junjie.sop@gmail.com

Z. Zou
e-mail: zouzq@njupt.edu.cn

X. He
e-mail: HXzcydyx@163.com

H. Dai
e-mail: haihongdai1@qq.com

Z. Zou
Jiangsu Key Laboratory of Big Data Security and Intelligent Processing,
Nanjing, Jiangsu 210023, China

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_5

1 Introduction

With the pervasiveness of the Global Positioning System (GPS) devices, a large number of GPS trajectories have been accumulating unobtrusively and continuously in daily life [1]. Since it could advance the recognition of individual behavioral patterns and facilitates the analysis of the crowd mobility and communication, visualizing and inferring the location of participants is one of the core tasks [2]. However, both recognizing human behavior and understanding a user's mobility from ST-data are critical issues in ubiquitous computing systems [3] and almost all of these systems still directly use raw GPS data, like coordinates and timestamps, without any processing and analysis. Hence, so far, these communities cannot provide much support for people with interesting information about geospatial locations. What's more, facing such a large dataset, users can hardly browse each GPS trajectory one by one. Now the Internet can provide a giant amount of information to a multitude of users, making Geographic Information System (GIS) service available to a wider range of public users than ever before. Web-based map services are the most important application of modern GIS systems. For example, Google Maps currently has almost one billion users. And the geo-enabled applications which use mobile devices provide web map services have a growing number at a great rate.

However, due to the highly complex and dynamic nature of GIS systems, end users can't quickly understand and effectively analyze ST-data, which is very challenging. First, the typical geographic visualization tools are complex and fussy with a lot of low-level details, making it difficult to use for ST-data analysis. Second, the analysis of large amounts of ST-data is very time costly. Third, current ST-data visualization tool is not well integrated into the map developer.

In this paper, to address the above challenges, we present the NUPT ST-data Miner, an Spatio-Temporal data analysis and visualization system, which is a type of intelligent processing system. By leveraging distributed computing, visualization, and ST-data mining, NUPT ST-data Miner enables users to perform visualizing specified user's mobile behavior, visualizing hot regions and support human decision intelligently. It also leverages rich user interactions to perform proposing possible actions and strategies and predicting the next position.

The remainder of this paper is organized as follows. In Sect. 2 discusses the related work. We describe the architecture and the system overview of NUPT ST-data Miner in Sect. 3. Section 4 we describe the system function and visualization in NUPT ST-data Miner GeoCloud. Section 5 studies the interfaces and the system performance. Finally, the conclusions and future work are laid out in Sect. 6.

2 Related Works

Many researchers have focused on real-time monitoring and event detection of the social media streams due to the new characteristics of GPS information, time stamps, texts, images, and videos embedded in the social media data and the prevalence of the various interesting real-world applications. Research topics covered in this area include event analysis, community detection, POI (point-of-interest) recommendation, prediction, and etc. [4–7]. For example, Nathan Eagle presented a complex system including the ethnographic studies of device usage, relationship inference, individual behavior modeling, and group behavior analysis [8]. Among all of these topics, POI recommendation and prediction are the most important topics due to the high value in both research and academy.

One line of the research, analyzing the social media streams, is conducted based on the news media data. Michael demonstrated an approach to achieve enhanced urban analysis by combining data from remote sensing and social media [9]. Yeon have presented a predictive visual analytics system using topic composition for text data, especially social media data and news media data, to forecast how text data for certain event evolve over time in the future [10]. StoryTracker shows visualizations of temporal flows for major topic groups as parallel color bars ranked by importance utilizing clustered news stream generated by the Europe Media Monitor (EMM) [11].

The other line of work focuses on the GPS information. Jiang focused on data-driven modeling method with trajectory data by cell phones. The influence of data noise in modeling process is discussed and feasibility of calibrating driving models with noisy data from cell phones is studied [12]. Bryan introduces epidemic disease simulation system to predicting spatial spread of the epidemic using agent-based models [13]. Schulz derives a visual exploration approach that consists of a novel multi-level visualization, adjoined traditional spatial and temporal views, as well as of tailored exploration techniques for their concerted use [14].

However, the visualization and analysis between the sequential successive ST-data doesn't exist a mature system yet. Although some work considers the characteristics of the ST-data, they mainly focus on explore and understand various patterns of topic trajectories or only for the time series analysis [15–17]. For example, He proposes a new approach to exploring the spatiotemporal text data with visual filters [16] but their functions are too simple and not easy to extend. TiMoVA's system only account for the time series analysis [15]. Comparatively speaking, our system could provide various kinds of data visualization and could be extended easily to add advanced functionality like machine learning.

3 System Overview

Figure 1 shows the framework of the NUPT ST-data Miner GeoCloud. In the data layer, the GPS trajectories with the location and timestamp would be available to us. Followed by pre-processing, users would make an analysis request and set parameters. After the analysis, they can visualize the user's stay point and the cluster on the map. In the soft layer, we will focus on using kinds of feature analysis module to fusion with the help of the parallel computing platform. In the hardware layer, firstly we use the small-scale multi machine construction of parallel computing environment to make the prediction, and then to the large-scale multicomputer deployed

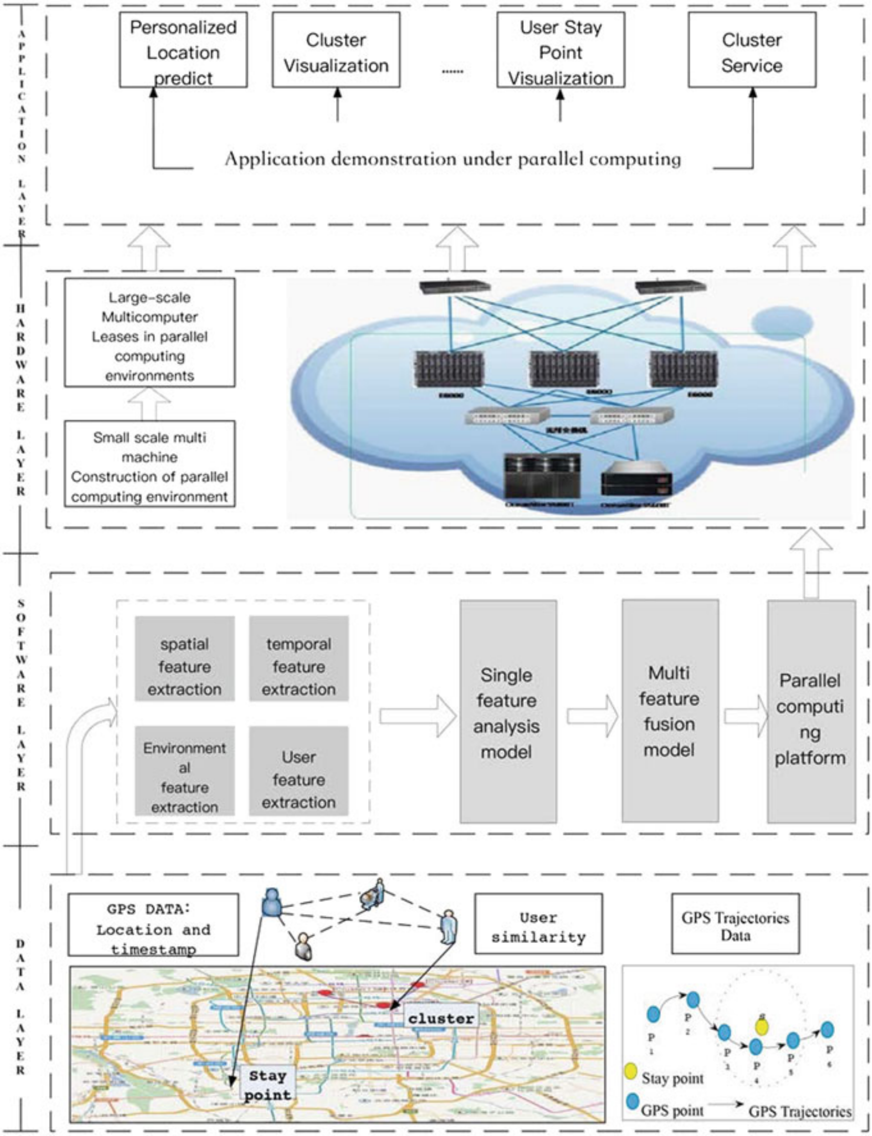
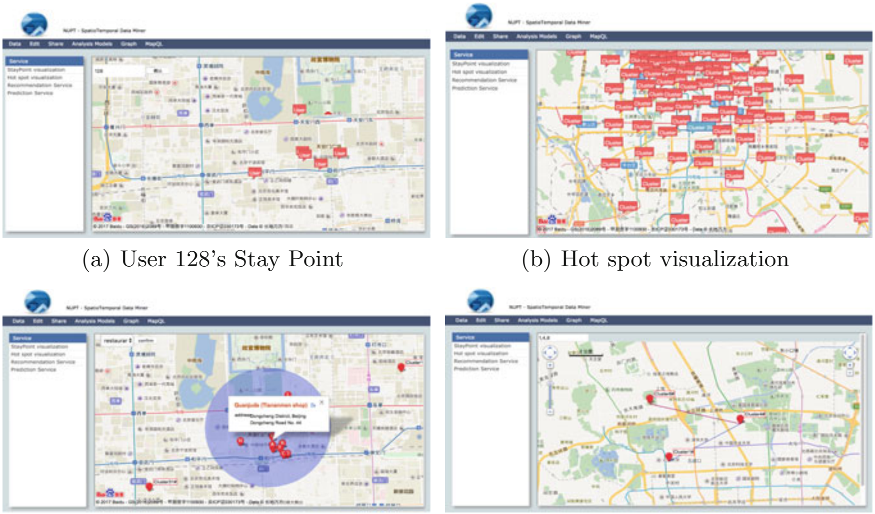

APPLICATION LAYER
Personalized Location predict
Cluster Visualization
......
User Stay Point Visualization
Cluster Service
Application demonstration under parallel computing
HARDWARE LAYER
Large-scale Multicomputer Leases in parallel computing environments
Small scale multi machine Construction of parallel computing environment
SOFTWARE LAYER
spatial feature extraction
temporal feature extraction
Environmental feature extraction
User feature extraction
Single feature analysis model
Multi feature fusion model
Parallel computing platform
DATA LAYER
GPS DATA: Location and timestamp
User similarity
GPS Trajectories Data
cluster
Stay point
P1
P2
P3
P4
P5
P6
Stay point
GPS point
GPS Trajectories

Fig. 1 The framework of NUPT data miner

in parallel computing environments, finally, we deploy our system to the Aliyun cloud. NUPT ST-data Miner Geocloud will provide the location prediction service, cluster visualization service, user stay points visualization service in our system's application layer.



(a) User 128's Stay Point

(b) Hot spot visualization

(c) Prediction visualization

(d) Recommendation service

Fig. 2 System function

4 System Function and Visualization

In this study, we demonstrate the visualization and the system functions. There are many ways to integrate ST-data mining and data visualization, however, the principle of our system visualization has four aspects, such as: stay point visualization, hot spot visualization, prediction service and recommendation service. The advantage of these is the mining results can be easily visualized. In addition, it has a good user interaction and could often be incorporated into the spatial mining process.

4.1 Stay Point Visualization

From each user's track log, we excavated out of the user's latitude, longitude and the timestamp. Firstly, we calculate the distance between a certain time interval by the latitude and longitude of the next point. Secondly, if the distance is short and the time interval is less than the preset estimate, we will fuse the two points close to the same time that they belong to the same stay point, and we calculate all the similar points to calculate. The center point is considered to be the stay point of the trajectory. As shown in Fig. 2a, we can use the keyword to retrieve all stay point within the current region of the specified user, in the center of the figure shows the user 128th, representing the 128th user's current region for the stay point.

4.2 Hot Spot Visualization

As stay point being extracted, you can clearly know all the user's meaningful track stop. Therefore, we aggregate all users with k-means algorithm. As shown in Fig. 2b, the cluster 31 is shown in the center of the graph, representing the 31st hot spot, and we present the hot spots after the clustering of all users matching the criteria to the user.

4.3 Prediction Service

The user's existing tracks are displayed in Fig. 2c. The system could predict the user's next point with the help of historical data. The blue line in the figure indicates the actual trajectory of the user, while the red line indicates the future trajectory of the user. When the user is about to reach the point at some time, we can advance the user service recommendations.

4.4 Recommendation Service

The system could predict the probability and frequency which the user reaches to the next point from the track in the frequent patterns. First of all, we circle around the center of the hot spot at a certain radius to the next point belongs. Then, if the user is a local citizen, system could recommend to the user the place according to the top-n possible choices, which are calculated by the history frequent log. If the user is a visitor, it is a cold start problem. However, we could recommend the highest frequency of local spots or the nearby businesses have promotional activities to him. As shown in Fig. 2d, the user is a local citizen; we search to get the peripheral services of the hot spot (such as restaurants, parking lots, shopping malls). In the center of the map shows the cluster 97, representing the 97th hot spot. If time is 11:30 am. We consider the distance and the best heat of Quanjude, and then the restaurant's telephone and address finally send to the user.

5 Visual Interfaces and System Performance

We have developed much system visualization function as shown in Fig. 2a–d. This system consists of several visual interactive components including stay point visualization, cluster visualization and prediction Service.

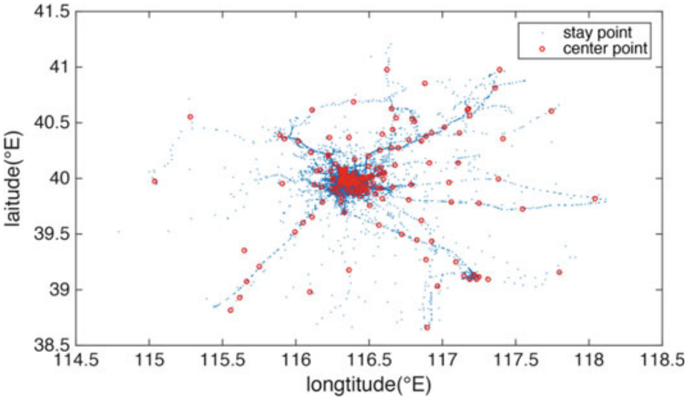


Fig. 3 All center points of 200 clusters and their stay point

5.1 Stay Point Visualization Interfaces

The purpose of the Stay Point Visualization Interfaces is to search for the user's stay points around the current region. To generate stay points from the dataset, we use the K-means algorithm [18] with two parameters to be T_{threh} 4 min and D_{threh} 200 m, and the results of generating stay point can be seen in Fig. 3 Based on the interface of the Baidu map, all the stay point could be displayed on the map.

5.2 Clustering Visualization Interfaces

In order to analysis a user's POI, too many stay point is not sufficient to represent the entire data since a stay point may not contain important information. Therefore, in this work, we provide Clustering Visualization Interfaces to generate clusters by different situations. To improve the speed of the clustering process, we utilized the Hadoop platform to run the K-means clustering algorithm. The result of the clustering stay points can be seen in Fig. 3. It depicts an experimental result with 200 clusters, in which the red points express the centers of clusters and the blue points express stay point.

5.3 Prediction Interfaces

Predictive Interfaces provides possible predictable positions in the future for the unknown event based on the analysis result. To develop our predictive interfaces, we have following three Steps to construct a combination of multiple characteristics based on mixture probabilistic model, which named ST-data Hybrid Prediction Model (STHPM).

First Step Combined Markov Model considers order-k sequential patterns with different k, and combines them together with smoothing which assigns weights on each pattern. Assuming the probability of the user moving to the position is $p_k(l_i)$.

$$p(l_i) = \sum_k p_k(l_i) * w_k, \tag{1}$$

$p(l_i)$ is the probability of location l_i with order-k context, and w_k is the weight of corresponding order-k context. Weights are predefined for ease of presentation [19].

Second Step Temporary model is assuming that the conditional probability of the user moving to the position at the next moment t_i is p_u^l, the corresponding action is C_i, the location of the action is l_i.

$$p_u^l(t_i|C_i = l_i) = p_u^l(t_i|\mu, \sigma^2) = \left(1/\sqrt{2\pi\sigma^2}\right)e^{-(t_i-\mu)^2/2\sigma^2} \tag{2}$$

$$\mu = (1/N)\sum_{i=1}^{N} t_i \tag{3}$$

$$\sigma^2 = (1/N)\sum_{i=1}^{N} (t_i - \mu)^2 \tag{4}$$

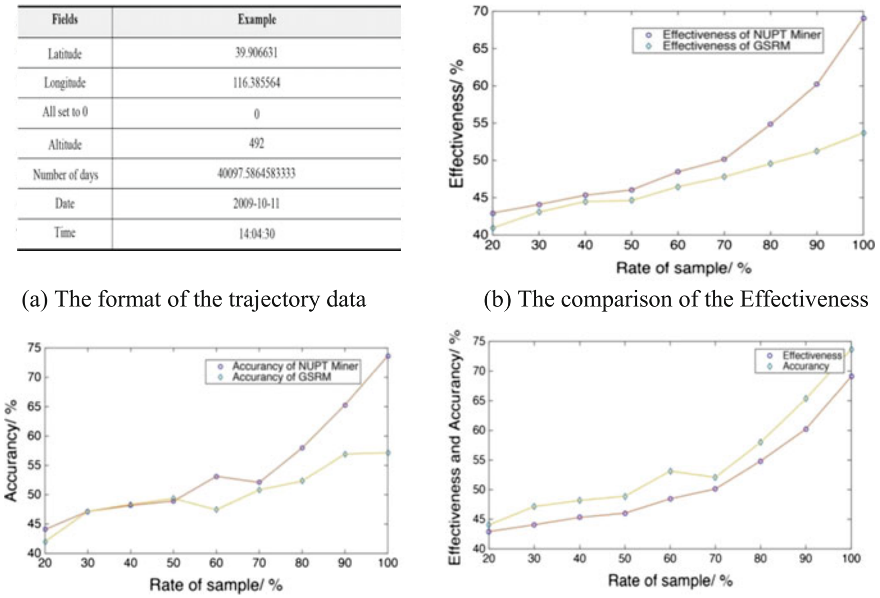
Final Step Our predictive interfaces is developed on the STHPM, which is a combination of the Combined Markov Model and Temporary model, establishes a probability-mixing model based on multiple features, and predicts the user's position at next moment. Assuming the probability that the user moves to a next point is $p_u(l)$.

$$p_u(l) = \sum_{k=1}^{2} \omega_k p_u^k(l) + \zeta_{u,l} \tag{5}$$

where k is the symbol of the feature model, ω_k is the weight coefficient, $p_u^k(l)$ is the result of the above calculation, $\zeta_{u,l}$ is the user u access to the location l's correction factor, the weight coefficient ω_k and the correction factor $\zeta_{u,l}$ are prepared by the machine learning method to get. Through the verified by multiple tests, finally we set ω_k to be the 0.12 and 0.88 to get an ideal prediction result.

5.4 Performance

Experiments were conducted based on real GPS trajectories from Microsoft Research Asia that included 182 users within a period of five years from April 2007 to August



Fields	Example
Latitude	39.906631
Longitude	116.385564
All set to 0	0
Altitude	492
Number of days	40097.5864583333
Date	2009-10-11
Time	14:04:30

(a) The format of the trajectory data

(b) The comparison of the Effectiveness

(c) The comparison of the Accuracy

(d) The influence of rate of samples

Fig. 4 The comparison between Our STHPM and GSRM

2012. In this dataset, a GPS trajectory is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude, and altitude, was measured every 1–5 s, which can be seen in Fig. 4a. The entire size of the GPS data for all users is 1.58 GB [18]. The clustering algorithm operates over a distributed environment that uses *Apache^{TM} Hadoop* (Hadoop Version 1.2.1) and is implemented in this prototype under Linux 12.04.

In order to further evaluate the predictions results produced by our model, two important metrics were defined, effectiveness and accuracy:

$$effectiveness = 100\% \times \frac{1}{x}\sum_{i=1}^{x}\frac{n_i}{m_i} \tag{6}$$

$$accuracy = 100\% \times \frac{1}{x}\sum_{i=1}^{x}\frac{t_i}{n_i} \tag{7}$$

where x is the total number of users, m_i is the total number of GPS trajectories of User i, n_i is the effective number of GPS trajectories among m_i, and t_i is the accurate predicted number of GPS trajectories among n_i.

We adopted the average rate of all prediction results during our experiments. The experimental results are shown in Fig. 4d. The trends of both curves are consistent with the theoretical analysis, the higher the number of samples, the higher the effectiveness and accuracy. From the experimental results shown in Fig. 4b, c, we can see that the two compares on the effectiveness and accuracy, which are from Our STHPM Method and the method of the Geographic Service Recommender Model (GSRM) [18].

The average accuracy of Our STHPM is 0.5447, better than the 0.4697 of the GSRM baseline. This is because our method exploits the historic user movement information, determines the internal relationship between spatial dimensions and temporal dimensions, and recommends locations with higher frequencies that have been visited.

We found that effectiveness and accuracy of predictions both decrease with a decreasing number of samples because the effectiveness of the mined Markov pattern decreases. Since we only used 182 users for our sample, which is likely not big enough. We plan to continue our research on this point in the future.

6 Conclusions and Future Work

In this paper we have presented an overview of NUPT ST-data Miner and introduced its interfaces, capabilities, and illustrate the various usage aspects of the system. Existing analysis workflows are broken into multiple loosely coupled which is suited for distributed computing steps. Since the analysis is deployed on the cloud computing platform, the load on network resources and local machine is minimized. Fast analytic functionalities are performed at the client side to further reduce the network load.

In the future, we would like to add more analysis algorithms to incorporate social factors in studying user behavior. We hope that this system will act as a building block for sophisticated visualization and analysis environments, tailored to suit the needs of many scientific communities.

Acknowledgements This work is supported by the National Natural Science Foundation of P. R. China (No. 41571389, 61472193, 41501431), supported by Key Laboratory of Spatial Data Mining Information Sharing of Ministry of Education, Fuzhou University (No. 2016LSDMIS07), and the NJUPT Natural Science Foundation (No. NY215116). In addition, we are grateful to the anonymous reviewers for their insightful and constructive suggestions.

References

1. Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from gps trajectories. In: International Conference on World Wide Web, pp 791–800
2. Wang F, Lu C-T, Qu Y, Philip SY (2017) Collective geographical embedding for geolocating social network users. In: Pacific-Asia conference on knowledge discovery and data mining, pp 599–611
3. Zheng Y, Li Q, Chen Y, Xie X, Ma WY (2008) Understanding mobility based on gps data. In: International Conference on Ubiquitous Computing, pp 312–321

4. Ertl T, Chae J, Maciejewski R, Bosch H, Thom D, Yun J, Ebert DS (2012) Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: IEEE conference on visual analytics science and technology, pp 143–152
5. Chandola V, Vatsavai RR, Bhaduri B (2011) Iglobe: an interactive visualization and analysis framework for geospatial data. In: International conference on computing for geospatial research applications, p 21
6. Nguyen H, Liu W, Rivera P, Chen F (2016) Trafficwatch: real-time traffic incident detection and monitoring using social media. In: Pacific-asia conference on knowledge discovery and data mining, pp 540–551
7. Wang X, Leckie C, Xie H, Vaithianathan T (2015) Discovering the impact of urban traffic interventions using contrast mining on vehicle trajectory data. In: Pacific-Asia conference on knowledge discovery and data mining, pp 486–497
8. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Personal Ubiquitous Computing 10(4):255–268
9. Jendryke M, Balz T, Mcclure SC, Liao M (2017) Putting people in the picture: combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in shanghai. Comput Environ Urban Syst 62:99–112
10. Yeon H, Yun J (2015) Predictive visual analytics using topic composition. In: International symposium on visual information communication and interaction, pp 1–8
11. Najm-Araghi M, Mansmann F, Krstaji M, Keim DA (2013) Story tracker: incremental visual text analytics of news story development. Inf Visual 12(3–4):308–323
12. Jiang Z, Yu S, Zhou M, Chen Y, Yi L (2017) Model study for intelligent transportation system with big data. Proced Comput Sci 107:418–426
13. Bryan C, Mniszewski S, Ma KL (2014) Integrating predictive visualization with the epidemic disease simulation system (episims). In: Proceedings of the IEEE VIS 2014 workshop visualization for predictive analytics
14. Schulz H-J, Hadlak S Schumann H (2013) A visualization approach forcross-level exploration of spatiotemporal data. In: International conference on knowledge management and knowledge technologies, p 2
15. Bo¨gl M, Aigner W, Filzmoser P, Gschwandtner T, Lammarsch T, Miksch S, Rind A (2014) Visual analytics methods to guide diagnostics for time series model predictions. In: Proceedings of the IEEE VIS 2014 workshop visualization for predictive analytics, VPA,
16. He J, Chen C (2016) Spatiotemporal analytics of topic trajectory. In: International symposium on visual information communication and interaction, 112–116
17. Bosch H, Thom D, Heimerl F, Puttmann E, Koch S, Kruger R, Worner M, Ertl T (2013) Scatterblogs2: real-time monitoring of microblog messages through user-guided filtering. IEEE Trans Vis Comput Graph 19(12):2022–2031
18. Zou Z, Yu Z, Cao K (2016) An innovative gps trajectory data based model for geographic recommendation service. Trans Gis
19. Gao H, Liu H (2015) Mining human mobility in location-based social networks. Synth Lect Data Min Knowl Discov 7(2):1–115

Applying Triple Data Encryption Algorithm to a Chaotic Systems: T-S Fuzzy Model-Based Approach

Feng-Hsiag Hsiao and Po-Han Lin

Abstract In the age of explosive growth in information exchanges, there is indeed for message security. Data Encryption Standard (DES) is one of the symmetric encryption algorithms which kept the dominant position in the area of data encryption over the last few decades. Nowadays, with a rapid development in the field of computer, the security of DES is too low when encountering the brute-force method for decryption. Therefore, 3DES applies the DES encryption algorithm three times to each data block to strengthen the complexity of the cryptosystem. Nevertheless, due to the meet-in-the-middle attack, it reduces the effective security of the ciphertext. Many studies have shown that most of the existing methods are unreliable in the aspect of security. Accordingly, this study proposes a double encryption using the 3DES and chaotic synchronization to increase the strength of the encryption. A design methodology for Takagi-Sugeno (T-S) fuzzy model-based secure communication in multiple time-delay chaotic (MTDC) systems is presented. The proposed encryption method produces a more secure communications system, while effectively protecting the encrypted message.

Keywords Triple data encryption algorithm · Block encryption · Improved genetic algorithm · Exponential synchronization · Chaotic synchronization

1 Introduction

Chaos is a well-known nonlinear phenomenon; it is a seemingly random event in a deterministic system characterized by sensitive dependence on initial conditions [1]. Because of these properties, chaos has interested scientists in various research fields [2, 3]. Chaotic synchronization has been extensively explored as a particular communication research field.

F.-H. Hsiao (✉) · P.-H. Lin
Department of Electrical Engineering, National University of Tainan, 33, Section 2,
Shu Lin Street, Tainan 700, Taiwan, ROC
e-mail: fhhsiao@mail.nutn.edu.tw

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_6

The chaotic synchronization proposed by Pecora and Carroll in 1990 [4] is intended to control one chaotic system to follow another. Based on this concept, diverse synchronization methods have been developed over the past two decades. Recently, many different control methods have attempted theoretically and experimentally to synchronize the chaotic systems. Such methods include adaptive control, observer-based control and fuzzy control, among others [5, 6]. To date, several types of fuzzy models have been developed and used in various technical and non-technical applications [7].

When analyzing the chaotic properties in observational time series, the problem of noise is unavoidable [8], and there will always be some noise or disturbances that may result in instability as external disturbances negatively affect the performance of chaotic systems. Accordingly, the motive of this paper is to achieve the exponential synchronization of multiple time-delay chaotic (MTDC) systems, and to simultaneously attenuate the influence of external disturbances on the control performance to a minimum level.

In recent years, the concept of an observer has been applied from a control theory perspective to the synchronization of chaotic systems [9]. A standard approach to solving the observer problem is to build a copy of the transmitter with added innovation that depends on the difference between the obtained signal and its prediction by the observer [10].

Data Encryption Standard (DES) was adopted in 1977 by the National Institute of Standards and Technology (NIST) [11, 12], which was grouped in 64-bit data encryption and decryption. And the data encryption and decryption algorithm are using the same structure, in which only the use of keys are in different order. The length of keys is 56-bit (the keys are usually expressed as 64-bit, but each eighth bit is used as parity check bit and can be ignored) [13]. Nowadays, the key length of DES is too short when using the brute-force method (a way to break the cipher by trying every possible key) for decryption, it only takes a few hours at a reasonable cost [14]. Therefore, Triple DES with three independent keys has a key length of 168 bits (three 56-bit DES keys) has taken the place of DES. However, because of the meet-in-the-middle attack [15], the effective security it provides is only 112 bits. This, results in a lower ciphertext security. In order to improve the strength of the ciphertext, this study carries out double encryption which combines chaotic synchronization with 3DES.

2 Problem Formulation

Consider two multiple time-delay chaotic (MTDC) systems in master-slave configuration. The dynamics of the master system (N_m) and slave system (N_s) are described as follows:

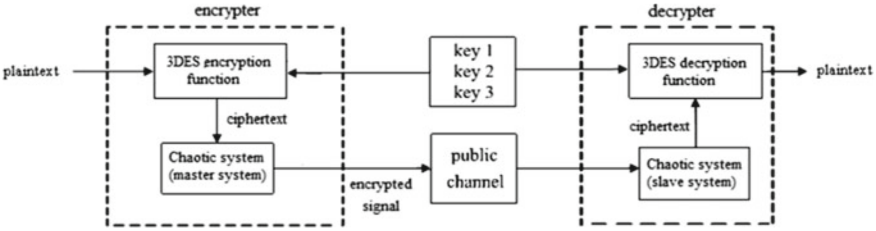


Fig. 1 Block diagram of the chaotic synchronization cryptosystem

$$N_m : \dot{X}(t) = f(X(t)) + \sum_{k=1}^{g} H_k(X(t - \tau_k)) \tag{2.1}$$

$$N_s : \dot{\hat{X}}(t) = f(\hat{X}(t)) + \sum_{k=1}^{g} H_k(\hat{X}(t - \tau_k)) + D(t) \tag{2.2}$$

where $f(\cdot)$, $\hat{f}(\cdot)$, $H_k(\cdot)$ and $\hat{H}_k(\cdot)$ are the nonlinear vector-valued functions, $\tau_k(k = 1, 2, \ldots, g)$ are the time delays and $D(t)$ denotes the external disturbance.

In this section, 3DES encryption function and three keys are first employed to encrypt the original message (plaintext) to produce the ciphertext, and it is re-encrypted via chaotic synchronization to realize the double encryption. A Takagi-Sugeno (T-S) fuzzy model is then established to approximate the MTDC system.

A chaotic synchronization cryptosystem is shown in Fig. 1. It consists of an encrypter (the master system and 3DES encryption function) and a decrypter (the slave system and 3DES decryption function). First, the encrypted message (ciphertext) are obtained by the plaintext and three keys (key1, key2, key3) via 3DES encryption function. The ciphertext is delivered to the master system and converted into the encrypted signal by chaotic masking The encrypted signal is then dispatched to the slave system through the public channel. Subsequently, the chaotic masking signal (encrypted signal) is filtered to obtain the ciphertext. Finally, the ciphertext can be decrypted by three keys and converted to the plaintext via 3DES decryption function.

2.1 3DES Cryptosystem

Data encryption standard (DES) is a symmetric encryption algorithm, it needs two inputs: a plaintext and a key. The length of the plaintext is 64 bits, and the key is also 64 bits in length (only 56 bits are ever used among them, the other 8 bits can be used as parity bits or simply set arbitrarily). DES is a typical block cipher which operates on blocks of data at a time. A block diagram of the algorithm is shown in Fig. 2.

However, the length of 56-bit is the disadvantage of DES. In order to enhance the strength of encryption, 3DES was designed to increase the key length of DES.

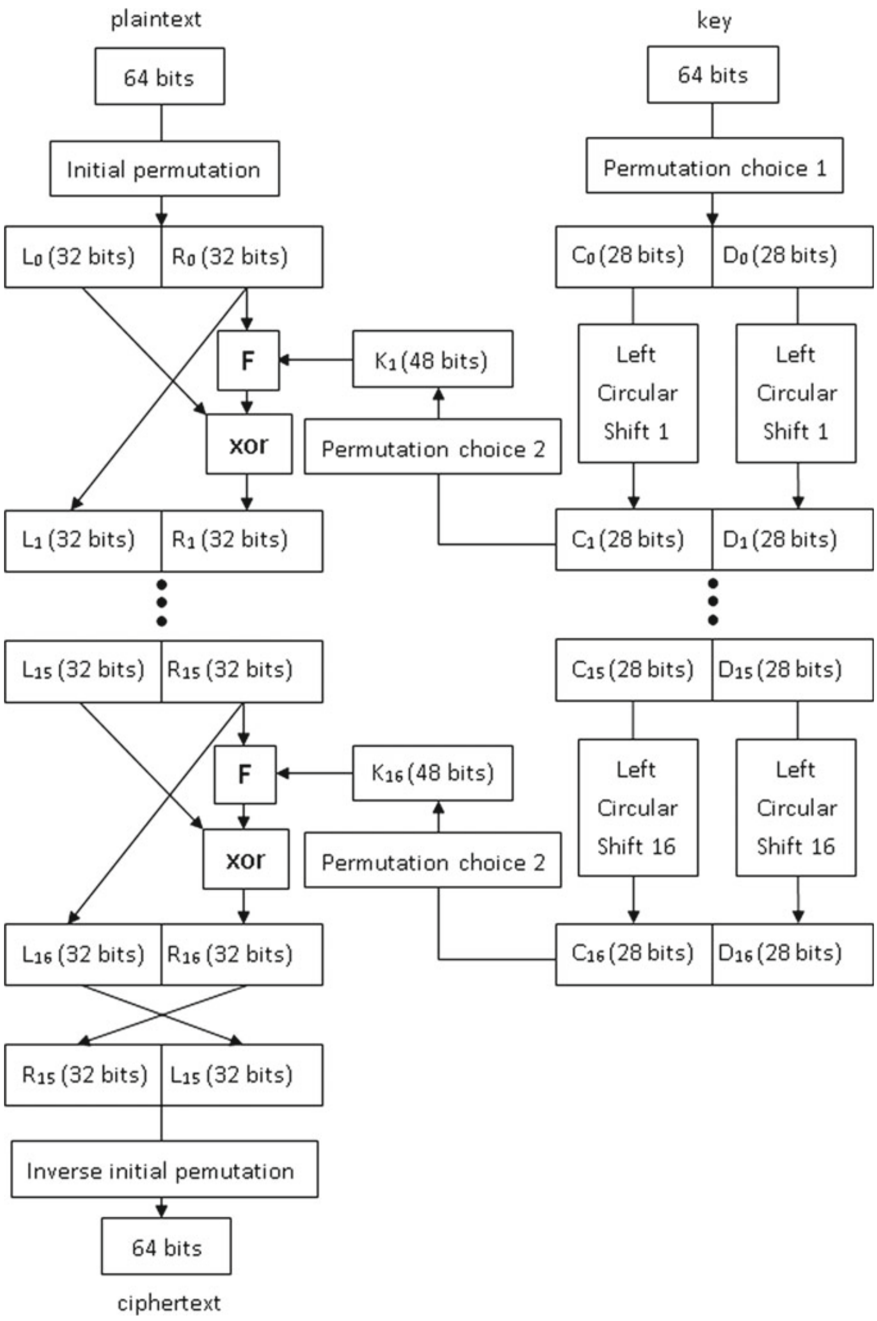


Fig. 2 The detailed structure of DES

3DES is also called EDE (Encrypt-Decrypt-Encrypt) which is a more secure version of DES, since it is used to encrypt the three DESs on the text so there are three different keys. 3DES has multiple mode of operations such as: Electronic Codebook (ECB), Cipher Block Chaining (CBC), Cipher Feedback (CFB), Output Feedback (OFB), Counter (CTR)…. In this study, we present CBC mode for 3DES encryption due to the advantage of using one key value for all three key inputs results in the same output as a single-DES encryption. The CBC mode uses a nonsecret 64-bit initializing vector (IV) and three 64-bit private keys ($\kappa_1, \kappa_2, \kappa_3$) to encrypt a plaintext *(P,P_2, …, P_n)* consisting of n 64-bit blocks to produce a ciphertext (Υ_1,Υ_2, …, Υ_n) consisting of n 64-bit blocks [11]. The 3DES encryption scheme is as below:

$$\Upsilon = E_{\kappa 3}(D_{\kappa 2}(E_{\kappa 1}(P))) \tag{2.3}$$

where P is plaintext, Υ is ciphertext, E_κ is conduct DES encryption with κ and D_κ is conduct DES decryption with κ.

The corresponding decryption scheme is as follows:

$$P = D_{\kappa 1}(E_{\kappa 2}(D_{\kappa 3}(\Upsilon))) \tag{2.4}$$

Therefore, the CBC mode of 3DES can be shown in Fig. 3.

2.2 The Takagi-Sugeno (T-S) Fuzzy Model

Over three decades ago, a fuzzy dynamic model was developed primarily from the pioneering work of Takagi and Sugeno [16] to represent the local linear input/output relations of nonlinear systems. This dynamical model is depicted by IF-THEN rules, and is used in this paper to process the synchronization problem of MTDC systems. The i th rule of the T-S fuzzy model for the master system is defined as follows:

Rule i: IF $x_1(t)$ is M_{i1} and … and $x_\delta(t)$ is $M_{i\delta}$

$$\text{THEN} \quad \dot{X}(t) = A_i X(t) + \sum_{k=1}^{g} \bar{A}_{ik} X(t - \tau_k)$$

where $i = 1, 2, \ldots, \phi$ and ϕ is the number of IF-THEN rules; A_i and $\bar{A}_{ik}$ are constant matrices with appropriate dimensions; $M_{i\eta}(\eta = 1, 2, \ldots, \delta)$ are the fuzzy sets, and $x_1(t) \sim x_\delta(t)$ are the premise variables. The final state of this fuzzy dynamic model is inferred as follows:

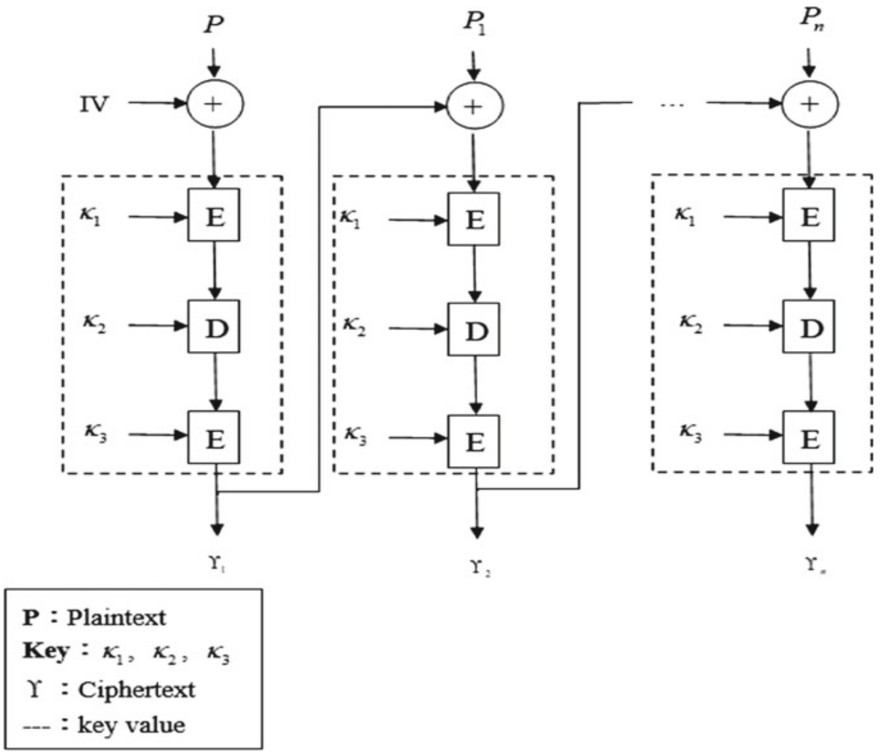


Fig. 3 Cipher Block Chaining mode of 3DES

$$\dot{X}(t) = \frac{\sum_{i=1}^{\phi} w_i(t)\left[A_i X(t) + \sum_{k=1}^{g} \bar{A}_{ik} X(t-\tau_k)\right]}{\sum_{i=1}^{\phi} w_i(t)}$$
$$= \sum_{i=1}^{\phi} h_i(t)\{A_i X(t) + \sum_{k=1}^{g} \bar{A}_{ik} X(t-\tau_k)\} \qquad (2.5)$$

where $w_i(t) \equiv \prod_{\eta=1}^{\delta} M_{i\eta}(x_\eta(t))$ and $M_{i\eta}(x_\eta(t))$ is the grade of membership of $x_\eta(t)$ in $M_{i\eta}$. Moreover, $h_i(t) \equiv \frac{w_i(t)}{\sum_{i=1}^{u} w_i(t)}$ and $\sum_{i=1}^{\phi} h_i(t) = 1$ for all t.

Similarly, the ℓ th rule of the T-S fuzzy model for the slave system is defined as follows:

Rule ℓ: IF $\hat{x}_1(t)$ is $\hat{M}_{\ell 1}$ and ... and $\hat{x}_\delta(t)$ is $\hat{M}_{\ell\delta}$

$$\text{THEN}\quad \dot{\hat{X}}(t) = A_\ell \hat{X}(t) + \sum_{k=1}^{g} \bar{A}_{\ell k}\hat{X}(t-\tau_k) + D(t)$$

where $\ell = 1, 2, \ldots, \sigma$ and σ is the number of IF-THEN rules; $\hat{A}_\ell$ and $\hat{\bar{A}}_{\ell k}$ are constant matrices with appropriate dimensions; $\hat{M}_{\ell\eta}(\eta = 1, 2, \ldots, \delta)$ are the fuzzy sets, and $\hat{x}_1(t) \sim \hat{x}_\delta(t)$ are the premise variables. The final state of this fuzzy dynamic model is inferred as follows:

$$\begin{aligned}\dot{\hat{X}}(t) &= \frac{\sum\limits_{\ell=1}^{\sigma} w_\ell(t)\left[A_\ell \hat{X}(t) + \sum\limits_{k=1}^{g} \bar{A}_{\ell k}\hat{X}(t-\tau_k) + D(t)\right]}{\sum\limits_{\ell=1}^{\sigma} w_\ell(t)} \\ &= \sum_{\ell=1}^{\sigma} h_\ell(t)\{A_\ell \hat{X}(t) + \sum_{k=1}^{g} \bar{A}_{\ell k}\hat{X}(t-\tau_k)\} + D(t)\end{aligned} \tag{2.6}$$

where $w_\ell(t) \equiv \prod\limits_{\eta=1}^{\delta} \hat{M}_{\ell\eta}(\hat{x}_\eta(t))$ and $\hat{M}_{\ell\eta}(\hat{x}_\eta(t))$ is the grade of membership of $\hat{x}_\eta(t)$ in $\hat{M}_{\ell\eta}$. In addition, $\hat{h}_\ell(t) \equiv \frac{w_\ell(t)}{\sum_{\ell=1}^{\sigma} w_\ell(t)}$ and $\sum_{\ell=1}^{\sigma} \hat{h}_\ell(t) = 1$ for all t.

2.3 Fuzzy Observer

For the fuzzy observer design, it is assumed that the dynamic fuzzy model of the slave system is observable. First, the fuzzy observers are designed based on the doublets (A_ℓ, C), as follows

Observer Rule ℓ: IF $\hat{x}_1(t)$ is $\hat{M}_{\ell 1}$ and … and $\hat{x}_\delta(t)$ is $\hat{M}_{\ell\delta}$

$$\begin{aligned}\text{THEN}\quad \dot{\hat{X}}(t) &= \hat{A}_\ell \hat{X}(t) + \sum_{k=1}^{g} \hat{\bar{A}}_{\ell k}\hat{X}(t-\tau_k) + Z_\ell(Y(t) - \hat{Y}(t)) + D(t) \\ \hat{Y}(t) &= C\hat{X}(t)\end{aligned}$$

where Z_ℓ is the observer gain, $\ell = 1, 2, \ldots, m$, where m is the number of IF-THEN fuzzy observer rules, and $M_{\ell\eta}(\eta = 1, 2, \ldots, \delta)$ are the fuzzy sets. *Y(t)* and $\hat{Y}(t)$ are the final outputs of the master system and the slave system, respectively. Therefore, the overall fuzzy observer is inferred as follows:

$$\begin{aligned}\dot{\hat{X}}(t) &= \sum_{\ell=1}^{m} \hat{h}_\ell(t)\{\hat{A}_\ell \hat{X}(t) + \sum_{k=1}^{g} \hat{\bar{A}}_{\ell k}\hat{X}(t-\tau_k)\} + Z_\ell(Y(t) - \hat{Y}(t)) + D(t) \\ \hat{Y}(t) &= C\hat{X}(t)\end{aligned} \tag{2.7}$$

In the past, solving the feedback gains, Z_ℓ ($\ell = 1, 2, \ldots, m$) was based on experience and trial-and-error. It would therefore be helpful to develop a powerful tool for solving suitable Z_ℓ ($\ell = 1, 2, \ldots, m$). This paper applies IGA to construct a novel algorithm for solving feedback gains [17].

2.4 Improved Genetic Algorithm

To improve the performance of the proposed method, this study adopts the IGA, whose superiority over standard GA. The key point of the IGA is that the chromosomes after crossover, on average, are arranged in the central and boundary regions of the search domain. This crossover gives the next generation more potential to find the global optimal solution.

The improved crossover is stated as follows [17]:

$$os_c^1 = [os_1^1\, os_2^1\, \ldots\, os_{no\_vars}^1] = \frac{P_1 + P_2}{2} \tag{2.8}$$

$$os_c^2 = [os_1^2\, os_2^2\, \ldots\, os_{no\_vars}^2] = P\text{max}(1 - w) + \max(P_1,\ P_2)w \tag{2.9}$$

$$os_c^3 = [os_1^3\, os_2^3\, \ldots\, os_{no\_vars}^3] = P_{\text{min}}(1 - w) + \min(P_1,\ P_2)w \tag{2.10}$$

$$os_c^4 = [os_1^4\, os_2^4\, \ldots\, os_{no\_vars}^4] = \frac{(P\text{max} + P_{\text{min}})(1 - w) + (P_1 + P_2)w}{2} \tag{2.11}$$

in which

$$P\text{max} = [para_{\text{max}}^1\, para_{\text{max}}^2\, \cdots\, para_{\text{max}}^{no\_vars}] \tag{2.12}$$

$$P_{\text{min}} = [para_{\text{min}}^1\, para_{\text{min}}^2\, \cdots\, para_{\text{min}}^{no\_vars}] \tag{2.13}$$

where $os_c^1 \sim os_c^4$ are the chromosomes of the next generation, P_1 and P_2 are the two chromosomes chosen from the parent, and $\max(P_1,\ P_2)$ and $\min(P_1,\ P_2)$ are the new chromosomes in which the genes are the maximum and minimum, respectively, of the genes in the two chromosomes P_1 and P_2. $para_{\text{max}}^{\vartheta}$, $para_{\text{min}}^{\vartheta}$ are the upper and lower bounds of the ϑ th genes, respectively, in the search space. Parameter $w \in [0,\ 1]$ is arbitrarily chosen. Equations (2.8) and (2.11) produce two new chromosomes distributed in the central region of the search domain, whereas (2.9) and (2.10) produce two new chromosomes distributed in the boundary region.

The fitness function for the application in this section is defined as follows:

$$Fit(\Lambda) = \frac{1}{1 + \sum_{t=0}^{t_f} \sum_{\eta=1}^{\delta} \left| e_\eta^\Lambda(t) \right|} \tag{2.14}$$

in which $Fit(\Lambda)$ is the fitness value of the Λth chromosome in a population, $e_\eta^\Lambda(t)$ is the error of the Λ th chromosome in a population.

The mutation operation serves to change the genes of the chromosomes. Consequently, the features of the chromosomes inherited from their parents can be changed [17]. Three new offspring will be generated by the mutation operation

$$nos_j = [os_1\ os_2\ \ldots\ os_{no\_vars}] + [b_1 \Delta os_1\ b_2 \Delta os_2 \ldots\ b_{no\_vars} \Delta os_{no\_vars}]\quad j = 1, 2, 3 \tag{2.15}$$

where b_i, $i = 1, 2, 3, \ldots, no\_vars$, can only take the value of 0 or 1, Δnos_i, $i = 1, 2, 3, \ldots, no\_vars$, are randomly generated numbers such that $para^i_{\mathrm{min}} \leq os_i + \Delta nos_i \leq para^i_{\mathrm{max}}$. These three new offspring will then be evaluated using the fitness function of (2.14). A real number will be generated randomly and compared with a user-defined number $p_a \in [\,0\ 1\,]$. If the real number is smaller than p_a, the chromosome with the largest fitness value of the three new offspring will replace the chromosome with the smallest fitness f_s in the population. If the real number is larger than p_a, the first offspring $\mathbf{nos_1}$ will replace the chromosome with the smallest fitness value f_s in the population if $f\ (\mathbf{nos_1}) > f_s$; the second and the third offspring will do the same. p_a is effectively the probability of accepting a bad offspring to reduce the chance of converging to a local optimum.

3 Stability Analysis and Chaotic Synchronization via Fuzzy Observer

In this section, the synchronization of multiple time-delay chaotic (MTDC) systems is examined under the influence of a modeling error. The exponential synchronization scheme of the MTDC systems is described below.

3.1 Master-Slave System

Based on the above section, the T-S fuzzy models of the master system with encrypted message (ciphertext) $\iota(\,\bullet\,)$ and the slave system under fuzzy observer are described as follows:

$$\begin{aligned}
\text{Master}:\quad & \dot{X}(t) = \sum_{i=1}^{u} h_i(t)\{A_i X(t) + \sum_{k=1}^{g} \bar{A}_{ik} X(t-\tau_k)\} + \iota(\,\bullet\,) \\
& Y(t) = CX(t) + \iota(\,\bullet\,) \\
\text{Slave}:\quad & \dot{\hat{X}}(t) = \sum_{\ell=1}^{m} \hat{h}_\ell(t)[\hat{A}_\ell \hat{X}(t) + \sum_{k=1}^{g} \hat{\bar{A}}_{\ell k} \hat{X}(t-\tau_k)] + Z_\ell(Y(t) - \hat{Y}(t)) + D(t) \\
& \hat{Y}(t) = C\hat{X}(t)
\end{aligned}$$

where $X(t)$ is the state vector and $Y(t)$ is the output vector of the master system. The state vector of the slave is $\hat{X}(t)$, and its output vector is $\hat{Y}(t)$. Z_ℓ is the observer gain, $D(t)$ is the external disturbance, and the encrypted message $\iota(\bullet)$ is inserted into the master systems.

3.2 Error Systems

From Eqs. (2.1) and (2.2), the synchronization error is defined as: $E(t) \equiv \hat{X}(t) - X(t) = [e_1(t),\ e_2(t),\ \ldots,\ e_\delta(t)].^T$ The dynamics of the error system under the fuzzy observer (2.10) can be described as follows:

$$\begin{aligned}\dot{E}(t) &= \hat{\Psi} + D(t) - \Psi \\ &= \sum_{i=1}^{u}\sum_{\ell=1}^{m} h_i(t)\left\{(A_i - Z_\ell C)E(t) + \sum_{k=1}^{g}\bar{A}_{ik}E(t-\tau_k)\right\} + D(t) + \Phi(t) \qquad (3.1)\end{aligned}$$

where $\hat{\Psi} \equiv \hat{f}(\hat{X}(t)) + \sum\limits_{k=1}^{g}\hat{H}_k(\hat{X}(t-\tau_k)) + Z_\ell(Y(t) - \hat{Y}(t))$, $\mathit{\Psi} = f(X(t)) + \sum\limits_{k=1}^{g} H_k(X(t-\tau_k)) + \iota(t)$ and $\Phi(t) \equiv \hat{\Psi} - \Psi - \left\{\sum\limits_{i=1}^{u}\sum\limits_{\ell=1}^{m} h_i(t)\left[(A_i - Z_\ell C)E(t) + \sum\limits_{k=1}^{g}\bar{A}_{ik}E(t-\tau_k)\right]\right\}$.

Suppose that there exists a bounding matrix $\varepsilon_{i\,l}^{qq}R$ such that:

$$\|\Phi(t)\| \leq \left\|\sum_{i=1}^{u}\sum_{\ell=1}^{m} h_i(t)\varepsilon_{i\,l}^{qq}RE(t)\right\| \qquad (3.2)$$

where R is the specified structured bounding matrix and $\|\varepsilon_{i\,l}\| \leq 1$, for $i = 1,\ 2,\ \ldots,\ u$; $l = 1,\ 2,\ \ldots,\ m$. Equations (3.2) show that:

$$\begin{aligned}\Phi^T(t)\Phi(t) &\leq \sum_{i=1}^{u}\sum_{\ell=1}^{m} h_i(t)\|RE(t)\|\left\|\varepsilon_{i\,\ell}^{qq}\right\|\sum_{i=1}^{u}\sum_{\ell=1}^{m} h_i(t)\left\|\varepsilon_{i\ell}^{qq}\right\|\|RE(t)\| \\ &\leq [RE(t)]^T[RE(t)] \qquad (3.3)\end{aligned}$$

Namely, $\Phi(t)$ is bounded by the specified structured bounding matrix R.

3.3 Delay-Dependent Stability Criterion for Exponential H$^\infty$ Synchronization

In this subsection, a delay-dependent criterion is proposed to guarantee the exponential stability of the error system described in (3.1). In addition, there will always be some noises or disturbances that may cause instability. External disturbance $D(t)$ negatively affects the performance of the chaotic systems. In order to reduce the effect of external disturbances, an optimal H^∞ scheme is used to design a fuzzy observer such that the effect of an external disturbance on the observer's performance can be attenuated to a minimum level. In other words, the fuzzy observer (2.7) simultaneously achieves exponential synchronization and realizes the optimal H^∞ control performance in this research.

Before examining the error system's stability, some definitions and lemma are given below.

Lemma 1 *[18]: For the real matrices A and B with appropriate dimension:*
$A^T B + B^T A \leq ƛ A^T A + ƛ^{-1} B^T B$
where ƛ is a positive constant.

Definition 1 [19]: The slave system (2.2) can exponentially synchronize with the master system (2.1) (i.e. the error system (3.1) is exponentially stable) if there exist two positive numbers α and β so that the synchronization error satisfies:

$$\|E(t)\| \leq \alpha\, exp(-\beta(t - t_0)), \quad \forall t \geq 0$$

where the positive number β is called the exponential convergence rate.

Definition 2 [9, 10]: The master system (2.1) and slave system (2.2) are in exponential H^∞ synchronization if the following conditions are satisfied:

(i) With zero disturbance (i.e. $D(t) = 0$), the error system (3.1) with the fuzzy observer (2.7) is exponentially stable.

(ii) Under zero initial conditions (i.e. $E(t) = 0$ for $t \in [-\tau_{\max}, 0]$, in which $\tau_{\max}$ is the maximal value of τ_k's) and a given constant $\rho > 0$, the following condition holds:

$$\Theta(E(t), D(t)) = \int_0^\infty E^T(t)E(t)dt - \rho^2 \int_0^\infty D^T(t)D(t)dt \leq 0, \tag{3.4}$$

where the parameter ρ is called the H^∞-norm bound or the disturbance attenuation level. If the minimum ρ is found to satisfy the above conditions (i.e. the error system can reject the external disturbance as strongly as possible), the fuzzy observer (2.7) is an optimal H^∞ synchronizer.

Theorem 1 *For given positive constants a, n and ξ, if there exists two symmetric positive definite matrices P and ψ_k, so that the following inequalities hold, then the exponential H^∞ synchronization with the disturbance attenuation ρ is guaranteed via the fuzzy observer (2.7):*

$$\begin{aligned}\Delta_{i\ell} &\equiv b\,(A_i - Z_\ell\, C)^T (A_i - Z_\ell\, C) + \sum_{k=1}^{g} \psi_k + n\,g\,R^T R \\ &\quad + I + \sum_{k=1}^{g} \tau_k^2 P^2 (b^{-1} + \xi^{-1} + n^{-1} + g\,a^{-1}) \qquad (3.5a)\\ \nabla_{i\,k} &\equiv ga\bar{A}_{ik}^T \bar{A}_{ik} - \psi_k \\ &< 0 \qquad (3.5b)\\ \rho &> \sqrt{\xi g} \qquad (3.5c)\end{aligned}$$

where $G_{i\,l} \equiv A_i - Z_l C$, for $i = 1,\ 2, \ldots,\ u$; $k = 1,\ 2,\ \ldots,\ g$ and $l = 1,\ 2,\ \ldots,\ m$ is the time delay, R is the specified structured bounding matrix shown in (3.2), and $\bar{A}_{i\,k}$ is described in (2.5).

By introducing the new variables, $Q = P^{-1}$, $F_l = Z_\ell Q$ and $\bar{\psi}_k = Q\psi_k Q^T$, According to Schur's complement [20], it is easy to show that the x inequalities in Eqs. (3.5a) and (3.5b) are equivalent to the following LMIs in Eqs. (3.6a) and (3.6b):

$$\begin{bmatrix} \Xi & QR^T & (A_i - Z_\ell C)Q^T \\ RQ^T & -(ng)^{-1}I & 0 \\ Q\,(A_i - Z_\ell C)^T & 0 & -(b^{-1})I \end{bmatrix} < 0 \qquad (3.6a)$$

$$\begin{bmatrix} -\bar{\psi}_k & Q\bar{A}_{i\,k}^T \\ \bar{A}_{i\,k}Q & -(ga)^{-1}I \end{bmatrix} < 0 \qquad (3.6b)$$

where

$$\Xi \equiv \sum_{k=1}^{g} \bar{\psi}_k + \sum_{k=1}^{g} \tau_k^2 (b^{-1} + \xi^{-1} + n^{-1} + ga^{-1})I + QIQ^T$$

Hence, Theorem 1 can be transformed into an LMI problem. Efficient interior-point algorithms are now available in the Matlab LMI Solver to solve this problem.

Corollary 1 *[21]: To verify the feasibility of solving the inequalities in Eqs. (3.6a) and (3.6b) using the LMI Solver (Matlab), interior-point optimization techniques are utilized to compute feasible solutions. These techniques require that the LMI systems are constrained to be strictly feasible, that is, the feasible set has a nonempty interior. For feasibility problems, the LMI Solver by feasp is shown as follows:*

$$\textit{Find} \quad x \quad \textit{such that the LMI} \quad L(x) < 0 \tag{3.7a}$$

as

$$\textit{Minimize} \quad t \quad \textit{subject to} \quad L(x) < t \times I \tag{3.7b}$$

where L(x) is a symmetric matrix and I is an identity matrix.

As mentioned above, the LMI constraint is always strictly feasible in x, t and the original LMI (3.7a) is feasible if and only if the global minimum $t$min of (3.7b) satisfies $t\text{min} < 0$. In other words, if $t\text{min} < 0$ will satisfy the inequalities (3.6a) and (3.6b), then the stability conditions (3.5a) and (3.5b) in Theorem 1 can be met. The obtained fuzzy observer (2.7) can then exponentially stabilize the error system; the H^{∞} control performance is realized at the same time.

Corollary 2 *In order to realize exponential optimal H^{∞} synchronization, the fuzzy observer design is formulated as the following constrained optimization problem:*

$$\textit{minimize}\ \rho > \sqrt{\xi g} \tag{3.8}$$

subject to $Q = Q^T > 0$, $\bar{\psi}_k = \bar{\psi}_k^T > 0$, (3.6a) and (3.6b).

More details on searching for the minimum ρ are given as follows:

The positive constant ξ is minimized by the mincx function of the Matlab LMI Toolbox. Accordingly, the minimum disturbance attenuation level $\rho_{\text{min}} > \sqrt{\xi_{\text{min}} g}$ can be obtained.

4 Conclusion

To prevent hackers from stealing personal information, this study propose double encryption systems via 3DES (Triple Data Encryption Algorithm) and chaotic synchronization. A design methodology for Takagi-Sugeno (T-S) fuzzy models-based secure communications in multiple time-delay chaotic (MTDC) systems to strengthen the complexity of the cryptosystem is presented. In addition, 3DES is used to execute triple DES encryption algorithms for each data block. However, due to the meet-in-the-middle attack, it reduces the effective security of the ciphertext. Accordingly, this research attempts to integrate chaotic synchronization with 3DES algorithm to increase the complexity of the cryptosystem. The proposed method can achieve a more secure communications system, while effectively protecting the encrypted message.

References

1. Hu C, Jiang H, Teng Z (2011) General impulsive control of chaotic systems based on a TS fuzzy model. Fuzzy Sets Syst 174:66–82
2. Poddar G, Chakrabarty K, Banerjee S (1998) Control of chaos in DC-DC converters. IEEE Trans. Circuit Syst I 45:672–676
3. Lin SL, Tung PC (2009) A new method for chaos control in communication systems. Chaos, Solitons Fractals 42:3234–3241
4. Pecora LM, Carroll TL (1990) Synchronization in chaotic systems. Phys Rev Lett 64:821–824
5. Mohammadzadeh A, Kaynak O, Teshnehlab M (2014) Two-mode indirect adaptive control approach for the synchronization of uncertain chaotic systems by the use of a hierarchical interval type-2 fuzzy neural network. IEEE Trans Fuzzy Syst 22:1301–1312
6. Lin TC, Lee TY (2011) Chaos synchronization of uncertain fractional-order chaotic systems with time delay based on adaptive fuzzy sliding mode control. IEEE Trans Fuzzy Syst 19:623–635
7. Jafari R, Yu W, Li X (2016) Fuzzy differential equations for nonlinear system modeling with bernstein neural networks. IEEE Access 4:9428–9436
8. Wang W, Gelder PHAJMV, Vrijling JK (2008) The effects of dynamical noises on the identification of chaotic systems: with application to streamflow processes. In: Fourth International Conference on Natural Computation, Jinan, 2008, pp 685–691
9. Li S, Xu W, Li R (2007) Synchronization of two different chaotic systems with unknown parameters. Phys Lett A 361:98–102
10. Parlitz U, Kocarev L, Schuster HG (1999) Handbook of chaos control. Wiley–VCH
11. Coppersmith D, Johnson DB, Matyas SM (1996) A proposed mode for triple-DES encryption. IBM J Res Dev 40:253–262
12. Mitchell CJ (2016) On the security of 2-key triple DES. IEEE Trans Inf Theory 62:6260–6267
13. Jun Y, Na L, Jun D (2009) A design and implementation of high-speed 3DES algorithm system. In: Second international conference on future information technology and management engineering, Sanya 2009, pp 175–178
14. Ren Y et al (2016) Key recovery against 3DES in CPU smart card based on improved correlation power analysis. Tsinghua Sci Technol 21:210–220
15. Handschuh H, Prenee B (1999) On the security of double and 2-key triple modes of operation. In: Fast software encryption, vol. 1636 of lecture notes in computer science, pp 215–230
16. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 15:116–132
17. Pan ST (2011) Evolutionary computation on programmable robust IIR filter pole-placement design. IEEE Trans Instrum Meas 60:1469–1479
18. Wang WJ, Cheng CF (1992) Stabilising controller and observer synthesis for uncertain large-scale systems by the Riccati equation approach. IEE Proc D 139:72–78
19. Sun YJ (2009) Exponential synchronization between two classes of chaotic systems. Chaos, Solitons Fractals 39:2363–2368
20. Limanond S, Si J (1998) Neural-network-based control design: an LMI approach. IEEE Trans Neural Netw 9:1422–1429
21. Gahinet P, Nemirovski A, Laub AJ, Chilali M (1995) LMI control toolbox user's guide. The MathWorks, Inc

A Framework for Performance Analysis of Various Load Balancing Techniques in a Software-Defined Networking Environment

Patrick Von Angelo V. Atienza and William Emmanuel S. Yu

Abstract Load balancer is an essential part of a computer network. Its primary purpose is to distribute incoming traffic across multiple target servers. There are numerous load balancing techniques and each of them excels on specific network topology and server capability. However, due to vendor dependency, implementing a quintessential load balancer requires additional hardware cost and knowledge in vendor-specific configurations. Using software-defined networking (SDN) approach, testing of various load balancing techniques becomes easier and cheaper than traditional hardware-based approach. Despite the promising advantages of SDN, the novel approach is still unstable. Hence, in this experiment, performances of five different load balancing techniques—namely, random, round-robin (RR), weighted round-robin (WRR), least-connections (LC), and weighted least-connections (WLC)—were tested. The experiment was done on a single-switch topology. Mininet and POX controller were used to setup the network environment. The load balancers were also tested in two types of network conditions: with and without TCP SYN floods. After several iPerf tests, results in both network conditions indicated that RR and LC load balancers were both more than twice as fast as the one without load balancing implementation and moderately faster than random load balancer. LC and WLC were slightly faster than RR and WRR without SYN floods while RR and WRR were slightly faster with SYN floods. Future works, like testing the framework on other types of network topologies or low-level load balancing techniques, could strengthen the substantiation of stability of using SDN approach.

Keywords Software-defined networking · Load balancing · Mininet
POX controller · Openflow

P. V. A. V. Atienza (✉) · W. E. S. Yu
Ateneo de Manila University, Loyola Heights, Quezon City, Philippines
e-mail: patrick.atienza@obf.ateneo.edu

W. E. S. Yu
e-mail: wyu@ateneo.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_7

1 Introduction

The main goal of software-defined networking (SDN) is to separate the control layer from the infrastructure layer to make the control layer programmable by end-users. With this emerging architecture, the network can be manipulated programmatically without touching the physical devices [1]. It also offers flexible, fast and cost-effective solution to continuously changing business requirements [1]. In traditional networking, it is impossible to split the infrastructure layer from the control layer. Therefore, end-users had to rely on the vendor for software network configurations and additional features. SDN tries to change that dependence in networking to an open networking system [2]. In SDN, a centralized controller manages the network flows that passes through network switches. SDN is developed to handle large networks like WANs, cloud computing networks and virtual networks. With the growth in today's network, data loss and degradation are highly susceptible; thus, an efficient algorithm that can handle large amount of load is necessary [2].

Load balancing is a method to distribute workload across multiple servers or other resources to achieve maximized throughput, minimized latency, and overload avoidance [3]. Using multiple components with load balancing, instead of a single component, may increase reliability through redundancy. Load balancing is one of the initial steps of Quality of Service (QoS) networking. Depending on the specifications of each server, the load balancer can direct the flow of each request based on the type of request it receives. Studies, like in the research of Chato and Yu, successfully split network flows of HTTP and media streaming requests in an SDN network using various QoS mechanisms [4]. Load balancers can also be used as a first line of defense of DDoS attacks. It can reroute the network flow of all of the suspected packets to a single node or drop the packet completely. Guevara et al. successfully implemented the detection and dropping of suspected packet using an intrusion detection and prediction system over a software-defined network [5]. Load balancers are very handy with SDN because the network flow of each network switch can be guided programmatically by the control layer [3, 6]. In this research, we will compare five basic load balancing techniques, namely, round-robin, weighted round-robin, random, least-connections, and weighted least-connections. This research seeks to answer the following research questions:

1. What is the advantage of implementing a load balancer in the controller pane compare to forwarding the network flow to a single server?
2. Will each tested load balancing method can handle its stability in TCP SYN flood attacks? Which of the load balancing methods will perform the best?
3. Will all the load balancers perform well in terms of accuracy and throughput? Which of the load balancing methods will perform the best?

2 Theoretical Background

2.1 *Load Balancing Techniques*

Load balancers can be split up in two types: static load balancers and dynamic load balancers. In static load balancing, processes do not depend on the current state of the network and processes are assigned prior to the execution of the network. On the other hand, in dynamic load balancing, processes change from time to time and the system needs to be recalibrated in order to distribute load equally through servers. Below are the most common static and dynamic load balancing techniques:

Round-robin Load Balancing. RR is a static load balancing technique and it is one of the simplest methods for network flow distribution. Going down the list of servers in the group, the round-robin load balancer forwards a client request to each server turn by turn. When it reaches the end of the list, the load balancer returns back to the initial server and do the method again [3, 6]. The method of selecting a server can operate with the worst-case time complexity of O(1) since the details of the last selected server can be stored. **Weighted Round-robin Load Balancing.** In WRR, a weight is assigned to each server depending on its network capability, its traffic-handling capacity, or its power of processing the received data [8]. The higher the weight, the larger the proportion of client requests the server receives. WRR is also a static load balancing technique. This is useful for servers that have different specifications [3, 6]. This technique can also achieve the same worst-case time complexity as the ones in RR. **Least-Connections Load Balancing.** When a virtual server is configured to use LC, the load balancer selects the server with the fewest active connections. LC is a dynamic load balancing technique and one of the default methods in load balancing, because, in most circumstances, it provides the best performance [3]. The method of selecting a server has a worst-case time complexity of O(n) with n as the number of servers. The reason is that the load balancers should check all the active connections of each server. **Weighted Least-Connections Load Balancing.** WLC is almost the same as the non-weighted one with an exception all servers are being weighted depending on their network handling capabilities or their own performances of processing received packets. This technique can also achieve the same worst-case time complexity as the ones in LC.

2.2 *OpenFlow, Mininet, and POX Controller*

OpenFlow is one of the early standards of SDN. As defined in the paper of the McKeown et al., the main feature of OpenFlow is to have a full control of all data packets roaming around the network [7]. The movements of the data packets, or "flows" as commonly referred to it, are controlled through user-defined rules and protocols. The details of each flow entry can be recorded in a flow table which can be also controlled programmatically [7, 8]. Each flow entry has three fields: (1) the

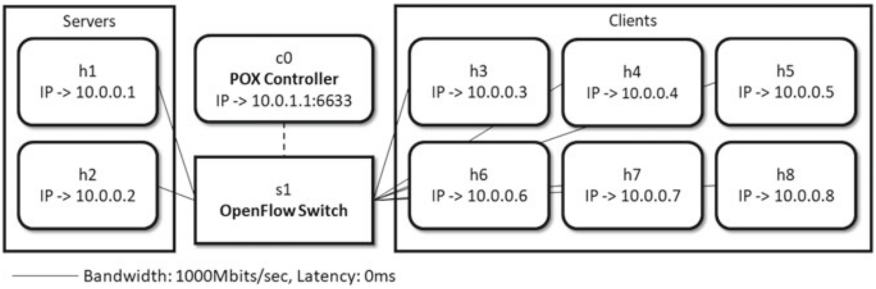


Fig. 1 The network topology to be set up using the Mininet network emulator

packet header that defines the flow or the “rule”, (2) the action that the packets are process or the “action”, and (3) the flow and port statistics or the “stats”. An action can be also classified into four basic types: (1) forwarding packet to ports, (2) forwarding and encapsulation of packets to controllers, (3) dropping of packets, and (4) forwarding packet to normal processing pipeline [7, 8].

Mininet is a network emulator that creates a virtual network of Layer 2 and Layer 3 switches, controllers, and hosts. The switches that offered by Mininet are Open vSwitches which support OpenFlow. Mininet is often used in research and development, prototyping, testing and debugging, and other tasks that needs an experimental network simulation. It runs on Unix/Linux environment [9].

POX controller is one of the most common frameworks for simulating SDN controllers. It is patterned from the NOX controller and it is written in Python. The main advantage of POX is that it is easy to use and does not require a steep learning curve [10, 11]. The disadvantage is that it is slow compare to other controllers like OpenDayLight, and Floodlight. That is why POX controllers are often used for educational purposes [10].

3 Methodology

3.1 Network Topology

The single-switch topology consisted of a single Open vSwitch (S1) connected to eight hosts (h1, h2, h3, h4, h5, h6, h7, h8) with two hosts (h1, h2) as servers and six hosts (h3, h4, h5, h6, h7, h8) as clients (Fig. 1). A switch was controlled by a remote POX controller. Each data link had a bandwidth of approximately 1000 megabits per second and 0 ms latency. An additional client host added in the network topology for the latter part of the experiment which tested the performance of the network during TCP SYN floods. The additional host acted as a persistent bot host.

3.2 Mininet and POX Controller Configurations

The network topology can be created programmatically in Python. However, Mininet has already terminal line commands on creating various types of networks based on network topologies such as single-switch topology.

```
sudo mn --topo single,8 --controller = remote,port = 6633 --link tc,bw = 1000,delay = 0 ms
```

The POX controller has an event called `PacketIn`. It triggers every time a packet is received by the controller. The packet can be identify as TCP if it returns a value in `packet.find('tcp')` command. If the source IP address of the TCP packet is a client IP address, the controller performs the load balancing method. On the other hand, if the source IP is a server IP address, then the controller forwards the TCP packet with the ACK response to the destination client. As explained by Peña and Yu, flow entries could be installed and modified by sending a `ofp_flow_mode()` message that matched the attributes of the packet [8].

As for the random load balancer, the controller picked a random IP address of a live server. The POX controller already had a configuration of random balancer which could be used anytime. As for the RR load balancer, the controller picked a live server with the lowest IP address in the array of live servers and saved the index of the server in the memory. Whenever a packet from a new TCP connection had been received, the controller would pick the succeeding live server and its IP address would be also saved. The process repeated after the controller picked the live server with the highest IP address. As for the LC load balancer, `FlowStatsReceived` event was used to determine active flows in the flow table. Each live server had a number of connections that could be incremented if it had the least value. All TCP connections established by client hosts had unique source port numbers. The controller saved that port number, with selected live server's IP address, to the memory. Using a polling function, the controller checked each port if it still existed in the flow stats. If the port number did not exist in the flow stats, the port number would be deleted in the memory and the corresponding live server's connections will be decremented. As for the WRR, the index will be set a float value and it will be incremented by 1 divided by the weight of previously picked server. As a basic rule of programming languages, if a float value is parsed to an integer, the result would be the floor value of the float variable. By these, using the integer-parsed value of the index would get the corresponding live server that can be picked by the controller. As for the WLC load balancer, the number of connections of each live server will be also set as a float value. Each connection that will be added to a live server must incremented to 1 divided the weight of the chosen live server.

4 Results

The Mininet-based network was tested on five load balancing algorithms. Each host transferred 1000 MB (or approximately 8388.608 megabits) of data to the server and this was done in iPerf 2.0.5 As for the weighted load balancers, h1 had a weight of 2 and host h2 would have a weight of 1; however, the host servers still retained the same specifications as the ones with non-weighted load balancers. Each load balancer was tested 10 times and the result of each test was averaged.

As shown in Table 1, The network with no implemented load balancer only uses an average of 87.53% of the total bandwidth. In contrast to the networks with non-weighted load balancers, the throughput of the non-load balanced network is extremely lower. The LC load balancer is more efficient than both RR and random load balancers with throughputs 1.53% and 11.25% higher respectively. Even in weighted load balancers, WLC was 1.79% faster and had a 0.85% higher throughput than WRR load balancer. LC and WLC were faster than RR and WRR in this scenario since the controller could pinpoint the best server for every initiated connection.

As for the part where TCP SYN floods were included, an additional host was added to continuously send 6 parallel TCP requests to the network. As shown in Table 2, the network without an implemented load balancer was significantly slower because all of the load of the TCP flood attacks was carried by only one host. The non-load balanced network had the least network throughput as it only used 39.85% of the total bandwidth. The RR load balancer had the fastest transfer time amongst all load balancers. RR and WRR were slightly better than the LC and WLC as they finished the whole transfer process 3% faster for non-weighted and 4.5% faster for weighted. One of the main reasons why LC and WLC were slower than RR and WRR in this scenario was because a single controller handled the decision making of all initiated connections including those that initiated by bot host. Since RR and WRR had better worst-cast time complexity than LC and WLC, the overhead of

Table 1 iPerf results of all load balancing techniques without TCP SYN flood

Load balancing method	Number of connections	Transfer time (s)	Transfer rate (Mbits/s)	% throughput
None	h1: 6, h2: –	h1: 57.5, h2: –	h1: 875.33, h2: –	h1: 87.53%, h2: –
Random	h1: 3, h2: 3	h1: 26.8, h2: 31.4	h1: 939.02. h2: 801.46	h1: 93.90%, h2: 80.15%
RR	h1: 3, h2: 3	h1: 25.9, h2: 26.9	h1: 971.65, h2: 935.53	h1: 97.17%, h2: 93.55%
LC	h1: 3, h2: 3	h1: 25.6, h2: 26.4	h1: 983.04, h2: 953.25	h1: 98.30%, h2: 95.33%
WRR	h1: 4, h2: 2	h1: 39.7, h2: 17.6	h1: 845.20, h2: 953.25	h1: 84.52%, h2: 95.32%
WLC	h1: 4, h2: 2	h1: 39.0, h2: 17.6	h1: 860.37, h2: 953.25	h1: 86.04%, h2: 95.32%

Table 2 iPerf results of all load balancing techniques with TCP SYN flood

Load balancing method	Number of connections	Transfer time (s)	Transfer rate (Mbits/s)	% throughput
None	h1: 6, h2: –	h1:126.3, h2: –	h1: 398.51, h2: –	h1: 39.85%, h2: –
Random	h1: 2.5, h2: 3.5	h1: 34.6, h2: 58.7	h1: 484.89, h2: 571.63	h1: 48.49%, h2: 57.16%
RR	h1: 3, h2: 3	h1: 45.4, h2: 53.3	h1: 554.31, h2: 472.15	h1: 55.43%, h2: 47.22%
LC	h1: 3, h2: 3	h1: 40.2, h2: 54.9	h1: 626.02, h2: 458.39	h1: 62.60%, h2: 45.84%
WRR	h1: 4, h2: 2	h1: 57.8, h2: 26.2	h1: 580.53, h2: 640.35	h1: 58.05%, h2: 64.04%
WLC	h1: 4, h2: 2	h1: 60.4, h2: 27.8	h1: 555.54, h2: 603.50	h1: 55.55%, h2: 60.35%

establishing each connection was less. It was evident that the penalty of the overhead was more prevalent than the benefit of the precise decision making of LC and WLC.

5 Conclusion and Future Works

The load balancers are successfully implemented in the POX controller. As the results show, implementing a load balancer within the control layer exceedingly increase the throughput of the network flow and decrease the transfer time by a hugely large amount. The results also show that the LC and RR load balancers are similarly efficient in terms of their throughput. However, a straightforward LC approach can be inefficient in terms of TCP requests because of the TIME_WAIT state of the TCP protocol that usually lasts around 2 min or less depending on its configuration. This can be problematic in a busy network and may possibly lead to an unstable load balancer.

This experiment is made to show the capability of an SDN controller to balance the requests of all clients. Most of the load balancing techniques are difficult to implement in a lower level network switched if SDN is not applied on the network topology. Future works like testing the performance of load balancers in other types of network topology or other types of low-level load balancing techniques, like Source IP Hashing and Least Packets, could strengthen the substantiation of the stability of using the software-defined networking approach. Implementing a testbed with a distributed controller system instead of a single controller could also avoid the bottleneck of the network.

References

1. Kreutz D, Ramos F, Verissimo P, Rothenburg C, Azodolmolky S, Uhlig S (2015) Software-defined networking: a comprehensive survey. Proc IEEE 103(1):14–76
2. Xia W, Wen Y, Foh CH, Niyato D, Xie H (2015) A survey of software-defined networking. IEEE Commun Surv Tutor 17(1):27–51
3. Bhandarkar S, Khan KA (2015) Load balancing in software-defined network (SDN) based on traffic volume. Adv Comput Sci Inf Technol (ACSIT) 2(7):72–76
4. Chato O, Yu W (2016) An exploration of various quality of service mechanisms in an OpenFlow and software defined networking environment in terms of latency and performance. In: 3rd international proceedings on information science and security (ICISS). IEEE, pp 1–7
5. Guevara AG, Domingo MA, Yu W (2017) Enhancing intrusion detection and prevention systems using software defined networking in a distributed topology. In: 17th proceedings on philippine computing science congress. CSP, Quezon City, Philippines, pp 219–228
6. Kaur S, Kumar K, Sing J, Ghumman N (2015) Round-robin based load balancing in software defined networking. In: 2nd international proceedings on computing for sustainable global development (INDIACom), IEEE, pp 2136–2139
7. OpenFlow Switch Specification. https://www.opennetworking.org/wp-content/uploads/2014/10/openflow-switch-v1.5.1.pdf. Accessed 17 Jan 2018
8. Peña JG, Yu W (2014) Development of a distributed firewall using software defined networking technology. In: 4th International Proceedings on Information Science and Technology (ICIST), IEEE, pp 449–452
9. Lantz B, Handigol N, Heller B, Jeyakumar V (2018) Introduction to Mininet. https://github.com/mininet/mininet/wiki/Introduction-to-Mininet. Accessed 17 Jan 2018
10. McCauley M (2018) POX Wiki. https://openflow.stanford.edu/display/ONL/POX+Wiki. Accessed 17 Jan 2018
11. Prete L, Shinoda A, Schweitzer C, de Oliveira R (2014) Simulation in an SDN network scenario using the POX controller. In: Proceedings on communications and computing (COLCOM). IEEE, pp 1–6

Inferring Social Relationships Through Network: A Systematic Literature Review

Fauqia Ilyas, Farooque Azam, Wasi Haider Butt and Kinza Zahra

Abstract Nowadays networks are developing extensively in size, intricacy, and diversity. Due to modification in social networks, advanced and distinctive kind of networks is emerging such as wireless networks, social networks, criminal networks and ego networks. Social network identification is the key to gather significant details from networks. Systematic Literature Review has been discerned to distinguish 31 papers from 2010 to 2018 to provide the set of frameworks that researchers could focus on. The aim is to organize the main categories of community discovery based on their definition of community and to identify algorithms, models, methods, and approaches that have been proposed. Consequently, 7 different categories of social networks have been identified. Furthermore, 20 algorithms, 4 approaches, 4 methods and 3 models for identifying social relationships from the network have been proposed. Based on the results obtained from the systematic review, we conclude that most of the work has been done on inferring community detection.

Keywords Social network detection · Community detection · Criminal networks Ego networks · Wireless networks

F. Ilyas (✉) · F. Azam · W. H. Butt · K. Zahra
Department of Computer Engineering, College of E&ME,
National University of Sciences and Technology (NUST),
H-12, Islamabad, Pakistan
e-mail: fauqia.ilyas85@ce.ceme.edu.pk

F. Azam
e-mail: farooq@ceme.nust.edu.pk

W. H. Butt
e-mail: wasi@ceme.nust.edu.pk

K. Zahra
e-mail: kinza.zahra15@ce.ceme.edu.pk

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_8

1 Introduction

The social network is a network of interactions and relationships portrays an association among social elements like companions, experts or co-creators. Through online social communities, there are plenty of chances for people to connect and team up with each other regardless of geographical locations. Each one of these systems is a huge database of a huge number of people and their actions [1]. Web-based social networking is one of the most common online communication devices [2].

Networks are an understandable process to illustrate communal, organic, mechanical and information frameworks. Junction in these networks arranges into densely associated categories that are normally alluded as a network circle, collections, and modules [3]. These frameworks are topic of current study venture.

Association extrication is the responsibility of recognizing connections between elements from unorganized to semi-organized information sources. Historically, papers on association extrication have been too great extent concentrated on recognizing already defined set of connections or distinguishing the statistical relationship between occurrences of already defined features. A mutual relationship is a short-term substantial relationship among two or more individuals. Some people together with an arrangement of connections among them contains a relational system, the research of such systems are affecting research community in many areas [4].

This paper aims to identify categories of social networks and to determine algorithms, models, approaches, and methods that have been proposed to infer social relationships. Hence, we attempt to obtain the appropriate answers to the following RQ's through systematic literature review.

RQ 1: What are the types of social networks so far has been addressed to identify relationships since 2010?
RQ 2: Which algorithms and models have been proposed to identify relationships from social networks?
RQ 3: Which approaches and methods have been suggested to identify relationships from social networks?

The remaining paper is systematically arranged as follows. Section 2 describes the methodology used in this review. Section 3 presents and discusses the review results. Section 4 provides the discussion and limitations. Section 5 presents an answer to RQ's. Conclusion and future work are presented in Sect. 6.

2 Methodology

In this study, we use systematic literature review [5] to accomplish this paper. We intend to determine algorithms, models, approaches, and methods that have been proposed to infer social relationships. Hence our research includes five phases: (1) Review Protocol development (2) Inclusion and Exclusion Criterion (3) Search Process (4) Quality Assessment (5) Data Extraction.

Table 1 Details of research works per database

Sr. #	Scientific database	Type	Selected research works	No. of researches
1	IEEE	Journal	[10, 11]	2
		Conference	[7, 12–16]	6
2	SPRINGER	Journal	[4]	1
		Conference	[6, 17, 18]	3
3	ELSEVIER	Journal	[1, 19, 20]	3
		Conference	[21]	1
4	ACM	Journal	[2, 3, 8, 22–27]	9
		Conference	[28–31]	4
5	Taylor and Francis	Journal	[9]	1
		Conference	[32]	1

2.1 Review Protocol Development

Inclusion and Exclusion Criteria

This study is carried out by following inclusion and exclusion rules:

1. We selected only those papers which dealt with inferring social relationships from the network.
2. We ensured the collection of latest studies by opting for those studies which lie in the years 2010 to 2018, and by not considering those studies which fall beyond the described period.
3. Primarily five leading scientific databases were used, which are IEEE, ACM, ELSEVIER, SPRINGER and TAYLOR and FRANCIS; to ensure the inclusion of authentic and state of the art research works. We opted for those papers which have been brought forward by the specified publishers. Details are given in Table 1.
4. We rejected redundant research studies and only most outstanding one of them was used.

Search Process

The search process comprises of the manual stage. To identify primary search references a manual search was applied. The initial stage was to select studies from databases like IEEE Explore, Elsevier, Springer, ACM, and Taylor and Francis. These digital libraries were picked as they were reviewed as the most significant and deliver excessive impact journals and conference proceedings that describe the areas of inferring social relationships from the network. From the formation and research questions of this review keyword that are used with the goal of finding as many related papers as possible accompanied by the intrinsic selection of keywords (Table 2). AND operator was used with appropriate terms. Figure 1 shows the steps performed during the search process.

Table 2 Details of search terms and search results

Sr. #	Search terms	Operator	IEEE	Springer	ACM	ELSEVIER	Taylor and Francis
1	Social network	AND	25649	219	11880	29094	1039
2	Social relationship detection	AND	346	3131	208	4696	28656
3	Relationship detection	AND	3706	8601	630	22237	1131
4	Network measurement	AND	60772	16881	7256	35273	972

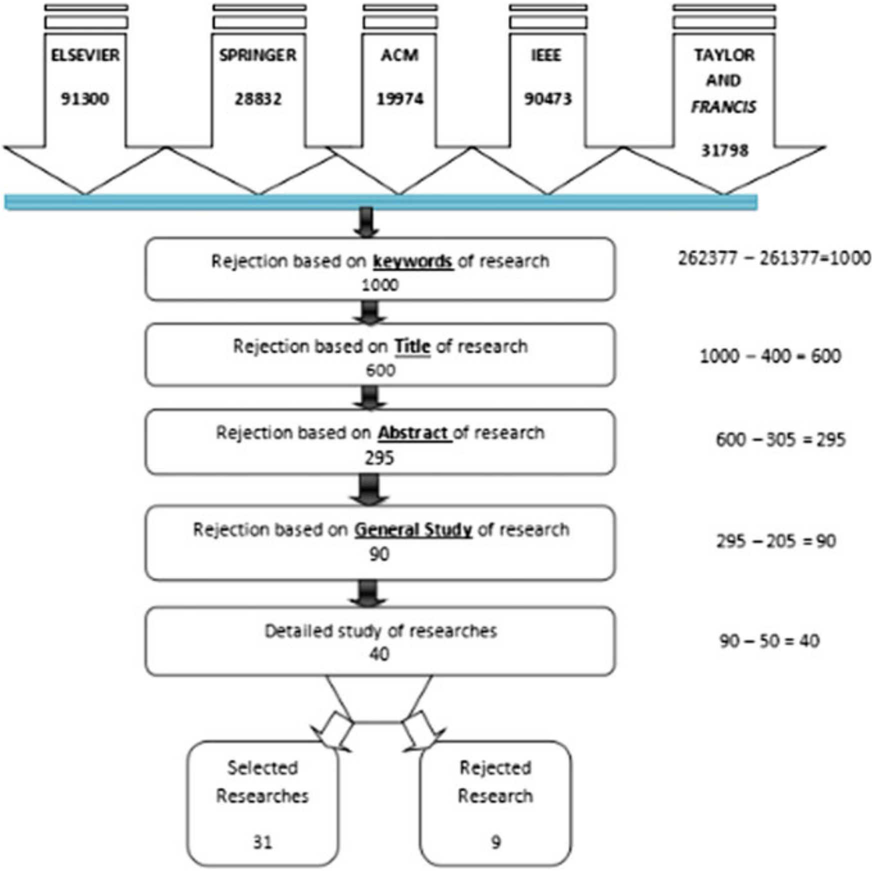


Fig. 1 Search process

Table 3 Data extraction and synthesis

Sr. #	Description	Details
1	Bibliographic information	Author, Title, Publication Year, publisher details, and type of research (i.e. journal or conference)
Extraction of data		
2	Overview	Aim of our selected study and what it is about
3	Results	Results taken from specified publications
4	Data Collection	Qualitative and quantitative method used
5	Assumptions	Assumption (if any) to authenticate the outcome
6	Validation	Validation of technique to authenticate its proposition
Synthesis of data		
7	Heterogeneity between researches distinguished	Recognition of diversity between papers
8	Illustration of the detection framework	Detection models and algorithms for inferring social relationships
9	Analytical implication and constancy for detecting social networks	Intensify the analytical implication for social relationship identification

Quality Assessment

To identify the quality of the network and model the analysis of results in the selected papers, these studies were evaluated through quality criteria.

QA1. Are the issues inscribed in the research linked with our analysis?
QA2. Is the context of the analysis explained in the study?
QA3. Is the method of review distinctly defined in the study?

Hence by the above quality assurance criteria we evaluated 31 studies in order to discover the plausibility of a certain selected study.

Data Extraction and Synthesis

Table 3 shows the data extraction and synthesis performed for our nominated researches to attain the answers of our RQ’s.

3 Results

The overview of chosen studies concerning digital libraries is illustrated in Table 1. This section presents and discusses the findings related to the systematic review questions. We screened 262377 records and included 31 publications that met our criteria. The selected studies were from 2010 to 2018 including descriptive studies using content analysis.

Table 4 Social network categories

Sr. #	Network categories	No. of researches	References
1	Co-Authorship network	2	[16, 21]
2	Community network	13	[2, 3, 6, 8–10, 14, 15, 19, 20, 25, 30]
3	Criminal network	4	[1, 17 18, 11]
4	Ego network	2	[12, 23]
5	Link prediction	2	[28, 31]
6	Overlapping community	6	[13, 24, 26, 27, 29, 21]
7	Wireless network	3	[4, 7, 22]

3.1 Social Network Categories

Accompanied by the development of the complex network in different fields, community detection has turned out to be an essential stage to comprehend the structure and dynamics of systems. Criminal networks are formed of criminals and association between these two. Another elementary, essential task of SNA is automatic detection in ego-networks. WSN is a wireless network composed of dimensional dispersed autonomous tools using sensors to monitor physical or environmental conditions. In collaborative author network, authors are the nodes and co-authorship of papers is the links. Identified categories of social networks are shown in Table 4.

3.2 Identified Algorithms and Model for Inferring Social Relationships from Network

Table 5 gives a detail of algorithms and models for identifying social relationships from network.

3.3 Identified Approaches and Methods for Inferring Social Relationships from Network

Table 6 gives a detail of approaches and methods for identifying social relationships from network.

Table 5 Identified algorithms and models for inferring social relationships from network

Algorithms		
Sr. #	Name of algorithms	References
Co-Authorship network		
1	Unique algorithm	[21]
Community network		
1	General algorithm	[2]
2	FCAN	[10]
3	MR-CPM	[15]
4	LOC	[25]
5	Novel community detection	[30]
6	Metaheuristic optimization	[19]
7	Divide and link	[20]
8	Discrete TL-GSO	[9]
Criminal network		
1	Adaptive detection algorithm	[18]
Ego network		
1	Novel circle detection	[12]
Link prediction		
1	FriendTNS	[28]
2	Supervised random walk	[31]
Overlapping community		
1	K-Clique percolation	[13]
2	Overlapping community detection	[24]
3	Node perception	[26]
4	Community detection	[27]
5	DEMON	[29]
6	Unique	[21]
Wireless network		
1	Hybrid clustering	[7]
2	Fuzzy C-means clustering	[22]
Models		
Sr. #	Name of models	References
Criminal network		
1	ComDM	[17]
2	Fuzzy SetBased AOM	[11]
Ego network		
1	Generative	[23]

Table 6 Identified approaches and methods for inferring social relationships from network

Approaches		
Sr. #	Name of approaches	References
Co-authorship network		
1	AuthorRank + FOAF	[16]
Community network		
1	Temporal coherence analysis	[6]
2	Incremental batch	[14]
Criminal network		
1	ADMOS	[1]
Methods		
Sr. #	Name of methods	References
Community detection		
1	Heuristic parameter-free community	[3]
2	CoDi	[8]
3	DISSECT	[32]
Wireless network		
1	Activity correlation spectroscopy	[4]

4 Discussion and Limitation

Communities in a network are the dense groups of the vertices, which are closely combined to each other inside the group and loosely combined to the rest of the vertices in the network. Community detection plays a fundamental part in understanding the functionality of complex networks. Identification of social structure can help us to understand network functionality. One of the vital tasks for social network analysis is by automatic social circle detection. The central user, the ego, is friends with all other users (the alters) in the network.

SNA can be applied to communications, community, complex networks, criminal networks and to detect structural hole. K-Clique percolation, Node Perception, and DEMON are the algorithms used for the identification of overlapping communities in the networks. Whereas several other algorithms such as Supervised Random Walks, LOC, MR-CPM and Metaheuristic optimization algorithms have previously been proposed to identify communities from the network. Consequently, CommDm and Fuzzy set based AOM and generative models are identified for inferring criminal and ego network respectively.

ADMOS, Incremental batch, Temporal Coherence analysis, and AuthorRank+ FOAF are the approaches that are used for detecting criminal, community and Co-Authorship networks. Whereas CoDi, DISSECT, and Heuristic parameter-free community are proposed methods used for community detection and Activity Correlation Spectroscopy for inferring dyadic social relationships using wireless networks.

Table 7 shows the comparison of algorithms, approaches, methods, and models that have been proposed for network detection based on the type of datasets used in publications. Most of the work has been done using real-world datasets (i.e., Facebook, Twitter, Google+, Amazon and mobile phone, etc.) whereas others category include datasets belong to intelligence, AlJihad, GN Benchmark, NetFlow, etc. Hybrid algorithms, approach, and method have been proposed for identifying social relationships [6, 7, 8, 9].

The main findings and the limitation of this review are discussed here. This review permits us to not only know about the state of the art and algorithms, but it also serves as a way to identify the principal contexts in which they were proposed and used while also giving an insight on the criteria used when facing the need to decide on what algorithms to use in different contexts and the existing algorithms that perform social relationship detection.

Limitations: However, we have attempted an exact search process to ensure the consistency and accuracy of researches, there is a possibility that some significant studies are missed.

5 Answer to Rq's

RQ 1: What are the types of social networks so far has been addressed to identify relationships since 2010?

Answer: 31 publications were carried out from 2010–2018 based on criteria mentioned in (Sect. 2). Through exhaustive study we have identified 7 different categories of networks that are used for social relationship identification as shown in Table 4.

RQ 2: Which algorithms and models have been proposed to identify relationships from social networks?

Answer: Through detailed analysis and screening of 31 publications since 2010 we have identified 20 algorithms and 3 models for inferring relationships from social networks as shown in Table 5.

RQ 3: Which approaches and methods have been proposed to identify relationships from social networks?

Answer: From selected 31 publications we have identified 4 approaches and 4 methods for identifying relationships from social networks as shown in Table 6.

Table 7 Comparative analysis of social relationship identification algorithms, models, approaches and methods

Network detection algorithms

Sr. #	Algorithms	Datasets		References
		Real world	Others	
1	General		✓	[2]
2	Adaptive detection	✓		[18]
3	FCAN	✓		[10]
4	Novel circle detection	✓		[12]
5	K-Clique percolation	✓		[13]
6	MR-CPM		✓	[15]
7	Hybrid clustering		✓	[7]
8	Fuzzy C-means clustering		✓	[22]
9	Overlapping community detection		✓	[24]
10	LOC		✓	[25]
11	Node perception	✓	✓	[26]
12	Community detection		✓	[27]
13	FriendTNS	✓		[28]
14	DEMON		✓	[29]
15	Novel community detection	✓		[30]
16	Supervised random walk	✓		[31]
17	Metaheuristic optimization	✓	✓	[19]
18	Divide and link		✓	[20]
19	Unique		✓	[21]
20	Discrete TL-GSO		✓	[9]

Network detection models

Sr. #	Models	Datasets		References
		Real world	Others	
1	ComDM		✓	[17]
2	Fuzzy SetBased AOM		✓	[11]
3	Generative	✓		[23]

Network detection approaches

Sr. #	Approaches	Datasets		References
		Real world	Others	
1	ADMOS		✓	[1]

(continued)

Table 7 (continued)

Network detection models

Sr. #	Models	Datasets		References
		Real world	Others	
2	Temporal coherence analysis		✓	[6]
3	Incremental batch	✓		[14]
4	AuthorRank + FOAF		✓	[16]

Network detection methods

Sr. #	Methods	Datasets		References
		Real world	Others	
1	Heuristic parameter-free community	✓	✓	[3]
2	Activity correlation spectroscopy	✓		[4]
3	CoDi	✓		[8]
4	DISSECT	✓		[32]

6 Conclusion and Future Work

This literature review examines various social network categories and set of frameworks for identifying social relationships from the network. SLR has been used to determine 31 researches published during 2010 to 2018. Following a review of inferring social relationships through the network, it is found that previous works are mostly based on community detection. Communities in the network are the collections of nodes, which are highly associated to each other than to the rest of nodes in the network. Type of network used as an input for community detection affects the use of resulting community outputs, e.g., identifying information flow in networks. Though acceptable efforts have been made for inferring social relationships from the network more scalable algorithms are required which reflect world community as results.

In future, it will be significant to infer community by lodging use of outside knowledge sources to extract more meaningful community and to consider the dynamic nature of the network to exploit dynamic communities.

References

1. Bindu PV, Thilagam PS, Ahuja D (2017) Discovering suspicious behavior in multilayer social networks. Comput Hum Behav 73:568–582
2. Liu T, Qin H (2016) Detecting and tagging users' social circles in social media. Multimedia Syst 22(4):423–431
3. Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213

4. Zhang X, Butts CT (2017) Activity correlation spectroscopy: a novel method for inferring social relationships from activity data. Soc Netw Anal Mining 7(1):1
5. Kitchenham B (2004) Procedures for performing systematic reviews, vol 33. Keele, UK, Keele University, pp 1–26
6. Tang X, Yang C, Gong X (2011) A spectral analysis approach for social media community detection. Soc Inf 127–134
7. Ferreira LN, Pinto AR, Zhao L (2012) QK-means: a clustering technique based on community detection and K-means for deployment of cluster head nodes. In: 2012 International Joint Conference on Neural Networks (IJCNN), June 2012, IEEE, pp 1–7
8. Ramezani M, Khodadadi A, Rabiee HR (2018) Community detection using diffusion information. ACM Trans Knowl Discov Data (TKDD) 12(2):20
9. Banati H, Arora N (2016) Detecting communities in complex networks-a discrete hybrid evolutionary approach. Int J Comput Appl 38(1):29–40
10. Hu L, Chan KC (2016) Fuzzy clustering in a complex network based on content relevance and link structures. IEEE Trans Fuzzy Syst 24(2):456–470
11. Shen Q, Boongoen T (2012) Fuzzy orders-of-magnitude-based link analysis for qualitative alias detection. IEEE Trans Knowl Data Eng 24(4): 649–664
12. Miao Q, Tang X, Quan Y, Deng K (2014) Detecting circles on ego network based on structure. In: 2014 Tenth International Conference on Computational Intelligence and Security (CIS), Nov 2014. IEEE, pp 213–217
13. Reid F, McDaid A, Hurley N (2012) August. Percolation computation in complex networks. In: 2012 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), Aug 2012. IEEE, pp 274–281
14. Chong WH, Teow LN (2013) An incremental batch technique for community detection. In: 2013 16th international conference on information fusion (FUSION), July 2013. IEEE, pp 750–757
15. Varamesh A, Akbari MK, Fereiduni M, Sharifian S, Bagheri A (2013) Distributed Clique Percolation based community detection on social networks using MapReduce. In: 2013 5th Conference on Information and Knowledge Technology (IKT), May 2013. IEEE, pp 478–483
16. Ahmedi L (2012) AuthorRank + FOAF: ranking for co-authorship networks on the web. In Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012) Aug 2012, IEEE Computer Society, pp 315–321
17. Ozgul F, Erdem Z, Bowerman C, Bondy J (2010) Combined detection model for criminal network detection. Intell Secur Inf 1–14
18. Xu L, Lin L, Wen S (2015) November. First-priority relation graph-based malicious users detection in mobile social networks. In: International conference on network and system security, Nov 2015. Springer International Publishing, pp 459–466
19. Atay Y, Koc I, Babaoglu I, Kodaz H (2017) Community detection from biological and social networks: a comparative analysis of metaheuristic algorithms. Appl Soft Comput 50:194–211
20. Gómez D, Zarrazola E, Yáñez J, Montero J (2015) A divide-and-link algorithm for hierarchical clustering in networks. Inf Sci 316:308–328
21. Li J, Wang X, Cui Y (2014) Uncovering the overlapping community structure of complexnetworks by maximal cliques. Phys A 415:398–406
22. Dutta R, Gupta S, Das MK (2014) Low-energy adaptive unequal clustering protocol using fuzzy c-means in wireless sensor networks. Wirel Pers Commun 79(2):1187–1209
23. Mcauley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Trans Knowl Discov Data (TKDD) 8(1):4
24. Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv (CSUR) 45(4):43
25. Wang R, Rho S, Cai W (2017) High-performance social networking: microblog community detection based on efficient interactive characteristic clustering. Clust Comput 1–13
26. Soundarajan S, Hopcroft JE (2015) Use of local group information to identify communities in networks. ACM Trans Knowl Discov Data (TKDD) 9(3):21

27. Jakalan A, Gong J, Su Q, Hu X, Abdelgder AM (2016) Social relationship discovery of IP addresses in the managed IP networks by observing traffic at network boundary. Comput Netw 100:12–27
28. Symeonidis P, Tiakas E, Manolopoulos Y (2010) Transitive node similarity for link prediction in social networks with positive and negative links. In: Proceedings of the fourth ACM conference on recommender systems Sept 2010. ACM, pp 183–190
29. Coscia M, Rossetti G, Giannotti F, Pedreschi D (2012) Demon: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining Aug 2012. ACM, pp 615–623
30. Dev H (2014) A user interaction based community detection algorithm for online social networks. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data, June 2014. ACM, pp 1607–1608
31. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on web search and data mining, Feb 2011. ACM, pp 635–644
32. Chin A, Chignell M, Wang H (2010) Tracking cohesive subgroups over time in inferred social networks. New Rev 16(1–2):113–139

Extended Flow Detection Scheme Using Simple Processor Trace (PT)

Hyuncheol Kim

Abstract A hardware tracer generates its enormous data into a log that is used for both performance analysis and debugging. Processor Trace (PT) is a new hardware-based tracing feature for Intel CPUs that traces branches executing on the CPU, which allows the reconstruction of the control flow of all executed code with minimal labor. Hardware tracer has been integrated into the operating system, which allows tight integration with its profiling and debugging mechanisms. In this paper, we implemented an in-line tracer that extends simple-pt in Linux. The in-line tracer proposed in this paper provides real-time information of symbol, specific address occurrence, frequency, overhead, and caller/callee based on information provided by simple-pt.

Keywords Tracing · Processor trace · Flow reconstruction · Flow detection

1 Introduction

For a long time, general processors have provided some dedicated hardware tracing modules to give some instrument to the developers when they desperately trying to fix bugs. This trend has been more stand out in the embedded systems category, however, rather than desktop or server-grade machines. For quite some time, embedded controllers have supported Joint Test Action Group (JTAG) interfaces that can be used to control debugging and tracing remotely. These specialized interfaces are usually connected with in-circuit emulators that utilize in-built debugging support in the processor hardware [1].

H. Kim (✉)
Department of Computer Science, Namseoul University, Cheonan, Korea
e-mail: hckim@nsu.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_9

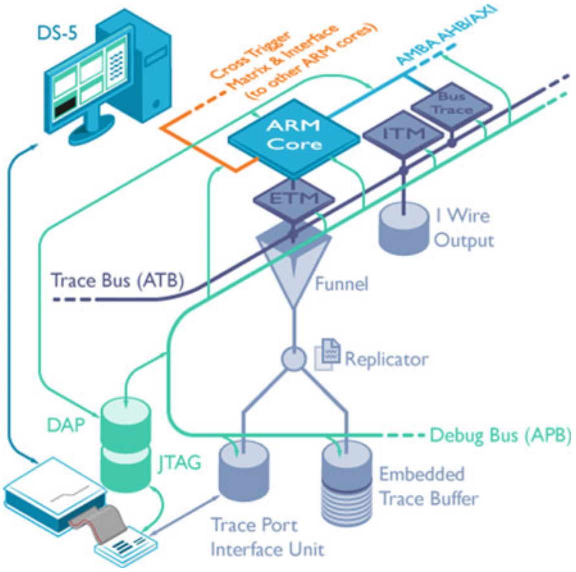


Fig. 1 Overview of ARM coresight

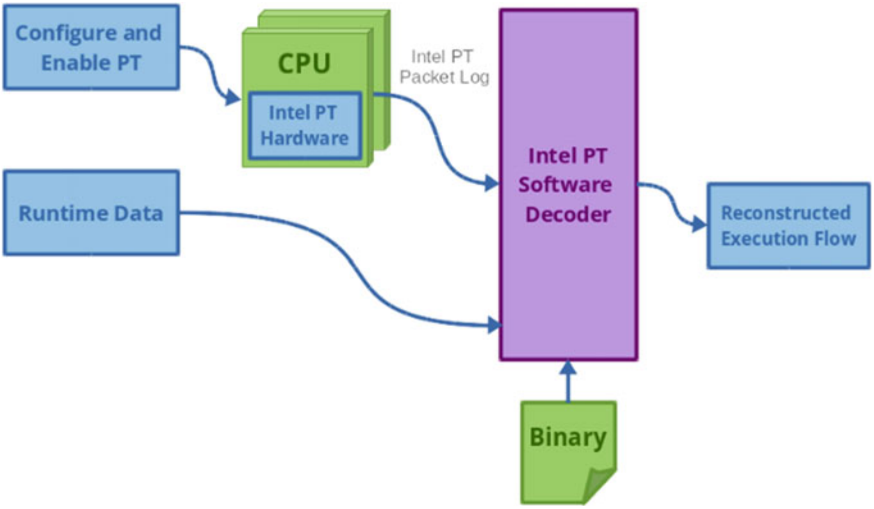
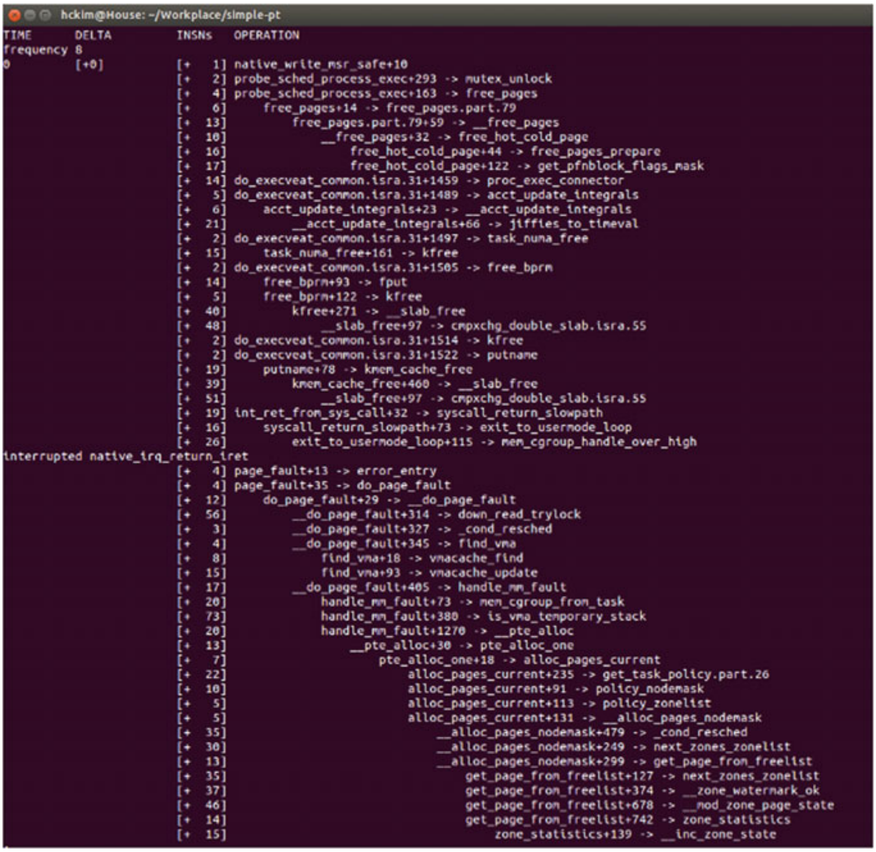


Fig. 2 Generic intel PT tracing flow

As shown in Figs. 1 and 2, typical hardware tracers include ARM Coresight and Intel PT. A hardware tracer generates its enormous data into a log that is used for both performance analysis and debugging. Processor Trace (PT) is a new hardware-based tracing feature for Intel CPUs that traces branches executing on the CPU, which allows the reconstruction of the control flow of all executed code with minimal labor.



```
TIME      DELTA         INSNs   OPERATION
frequency 8
0         [+0]          [+   1] native_write_msr_safe+10
                        [+   2] probe_sched_process_exec+293 -> mutex_unlock
                        [+   4] probe_sched_process_exec+163 -> free_pages
                        [+   6]     free_pages+14 -> free_pages.part.79
                        [+  13]         free_pages.part.79+59 -> __free_pages
                        [+  10]             __free_pages+32 -> free_hot_cold_page
                        [+  16]                 free_hot_cold_page+44 -> free_pages_prepare
                        [+  17]                 free_hot_cold_page+122 -> get_pfnblock_flags_mask
                        [+  14] do_execveat_common.isra.31+1459 -> proc_exec_connector
                        [+   5] do_execveat_common.isra.31+1489 -> acct_update_integrals
                        [+   6]     acct_update_integrals+23 -> __acct_update_integrals
                        [+  21]         __acct_update_integrals+66 -> jiffies_to_timeval
                        [+   2] do_execveat_common.isra.31+1497 -> task_numa_free
                        [+  15]     task_numa_free+161 -> kfree
                        [+   2] do_execveat_common.isra.31+1505 -> free_bprm
                        [+  14]     free_bprm+93 -> fput
                        [+   5]     free_bprm+122 -> kfree
                        [+  40]         kfree+271 -> __slab_free
                        [+  48]             __slab_free+97 -> cmpxchg_double_slab.isra.55
                        [+   2] do_execveat_common.isra.31+1514 -> kfree
                        [+   2] do_execveat_common.isra.31+1522 -> putname
                        [+  19]     putname+78 -> kmem_cache_free
                        [+  39]         kmem_cache_free+460 -> __slab_free
                        [+  51]             __slab_free+97 -> cmpxchg_double_slab.isra.55
                        [+  19] int_ret_from_sys_call+32 -> syscall_return_slowpath
                        [+  16]     syscall_return_slowpath+73 -> exit_to_usermode_loop
                        [+  26]         exit_to_usermode_loop+115 -> mem_cgroup_handle_over_high
interrupted native_irq_return_iret
                        [+   4] page_fault+13 -> error_entry
                        [+   4] page_fault+35 -> do_page_fault
                        [+  12]     do_page_fault+29 -> __do_page_fault
                        [+  56]         __do_page_fault+314 -> down_read_trylock
                        [+   3]         __do_page_fault+327 -> _cond_resched
                        [+   4]         __do_page_fault+345 -> find_vma
                        [+   8]             find_vma+18 -> vmacache_find
                        [+  15]             find_vma+93 -> vmacache_update
                        [+  17]         __do_page_fault+405 -> handle_mm_fault
                        [+  20]             handle_mm_fault+73 -> mem_cgroup_from_task
                        [+  73]             handle_mm_fault+380 -> is_vma_temporary_stack
                        [+  20]             handle_mm_fault+1270 -> __pte_alloc
                        [+  13]                 __pte_alloc+30 -> pte_alloc_one
                        [+   7]                     pte_alloc_one+18 -> alloc_pages_current
                        [+  22]                         alloc_pages_current+235 -> get_task_policy.part.26
                        [+  10]                         alloc_pages_current+91 -> policy_nodemask
                        [+   5]                         alloc_pages_current+113 -> policy_zonelist
                        [+   5]                         alloc_pages_current+131 -> __alloc_pages_nodemask
                        [+  35]                             __alloc_pages_nodemask+479 -> _cond_resched
                        [+  30]                             __alloc_pages_nodemask+249 -> next_zones_zonelist
                        [+  13]                             __alloc_pages_nodemask+299 -> get_page_from_freelist
                        [+  35]                                 get_page_from_freelist+127 -> next_zones_zonelist
                        [+  37]                                 get_page_from_freelist+374 -> __zone_watermark_ok
                        [+  46]                                 get_page_from_freelist+678 -> __mod_zone_page_state
                        [+  14]                                 get_page_from_freelist+742 -> zone_statistics
                        [+  15]                                     zone_statistics+139 -> __inc_zone_state
```

Fig. 3 Example of simple-PT sptdecode command

PT can lookout the instructions of the user and kernel-level code, and can collect cycle information up to the detailed steps of instruction branches. Compared to the existing tracing information, Processor Trace is much faster and more flexible in terms of what type and amount of trace information can be recorded. This allows the construction of very detailed execution flows and facilitates performance and correctness debugging at the level of maximum precision [2].

In the meantime, hardware tracer has been integrated into the operating system, which allows tight integration with its profiling and debugging mechanisms. This makes it possible to use different trace buffers for different processes, and to make the facility available for non-root users [3, 4]. simple-pt is a simple implementation of Intel PT on Linux [4, 5]. simple-pt decodes the branch trace and displays a function or instruction level trace. In this paper, we implemented an in-line tracer that extends simple-pt in Linux. The in-line tracer proposed in this paper provides real-time information of symbol, specific address occurrence, frequency, overhead, and caller/callee based on information provided by simple-pt.

2 Simple Processor Trace (Simple-PT)

2.1 Overview

Simple-pt is a simple implementation of Intel Processor Trace (PT) on Linux. PT can trace all branches executed by the CPU at the hardware level with moderate overhead [5, 6, 7]. simple-pt consists of (1) kernel driver, (2) sptcmd to collect data from the kernel driver, (3) sptdecode to display function or instruction traces, and (4) fastdecode to dump raw PT traces [6]. Simple-pt uses the libipt PT decoding library. sptcmd loads and configures the kernel driver. sptcmd traces while running the program and always performs global tracing. sptcmd writes PT trace data to each trace file (ptout.N: N is CPU #) for each CPU. It also records the sideband information needed for decode in the ptout.sideband file. As shown in Fig. 3, sptdecode decodes the CPU trace information using sideband information. To decode kernel code, you must be root to read/proc/kcore.

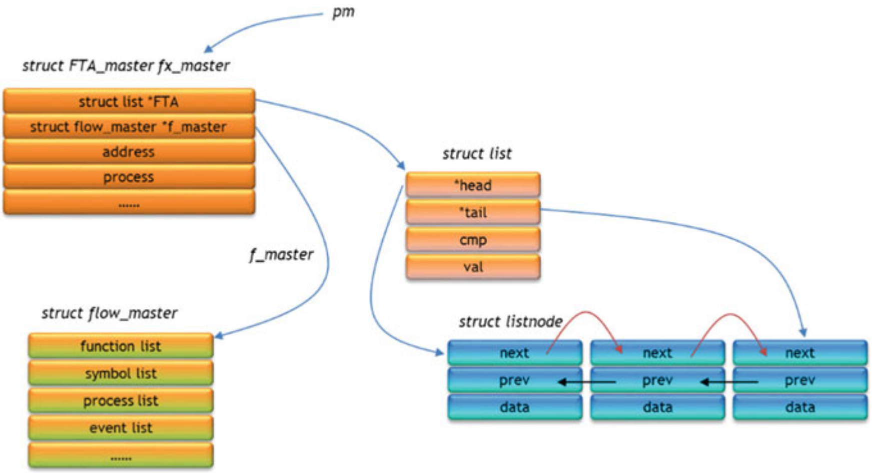
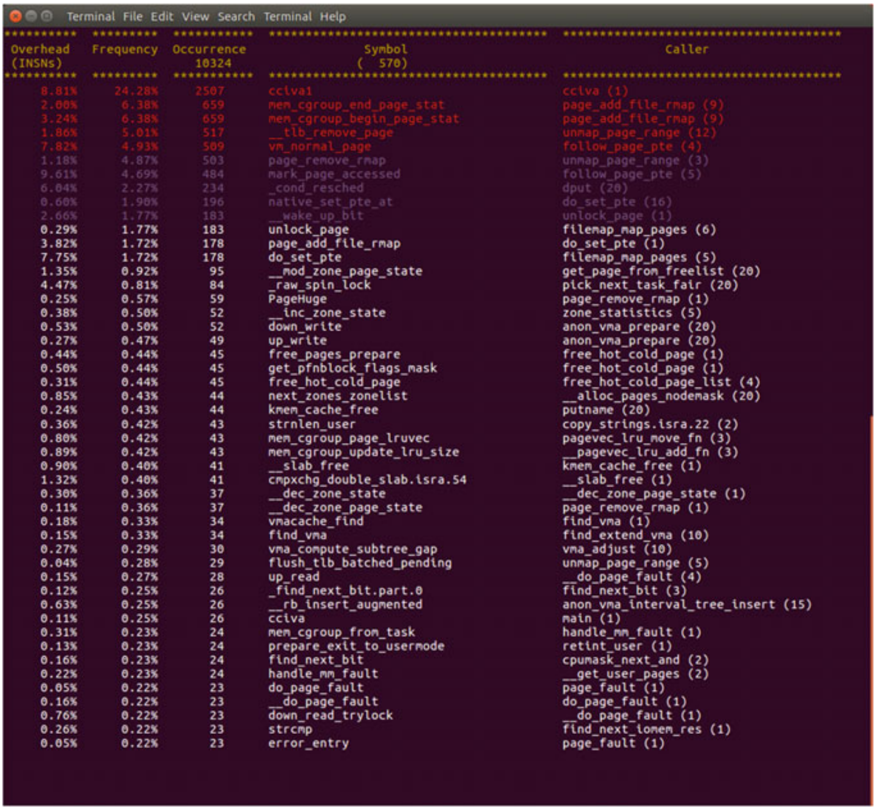


Fig. 4 Data structure for flow trace analyzer



```
**********  *********  ***********  **************************************  **************************************
Overhead    Frequency  Occurrence                  Symbol                                  Caller
(INSNs)                  10324                    (  570)
**********  *********  ***********  **************************************  **************************************
     8.81%     24.28%      2507     cclva1                                  cclva (1)
     2.00%      6.38%       659     mem_cgroup_end_page_stat                page_add_file_rmap (9)
     3.24%      6.38%       659     mem_cgroup_begin_page_stat              page_add_file_rmap (9)
     1.86%      5.01%       517     __tlb_remove_page                       unmap_page_range (12)
     7.82%      4.93%       509     vm_normal_page                          follow_page_pte (4)
     1.18%      4.87%       503     page_remove_rmap                        unmap_page_range (3)
     9.61%      4.69%       484     mark_page_accessed                      follow_page_pte (5)
     6.04%      2.27%       234     _cond_resched                           dput (20)
     0.60%      1.90%       196     native_set_pte_at                       do_set_pte (16)
     2.66%      1.77%       183     __wake_up_bit                           unlock_page (1)
     0.29%      1.77%       183     unlock_page                             filemap_map_pages (6)
     3.82%      1.72%       178     page_add_file_rmap                      do_set_pte (1)
     7.75%      1.72%       178     do_set_pte                              filemap_map_pages (5)
     1.35%      0.92%        95     __mod_zone_page_state                   get_page_from_freelist (20)
     4.47%      0.81%        84     _raw_spin_lock                          pick_next_task_fair (20)
     0.25%      0.57%        59     PageHuge                                page_remove_rmap (1)
     0.38%      0.50%        52     __inc_zone_state                        zone_statistics (5)
     0.53%      0.50%        52     down_write                              anon_vma_prepare (20)
     0.27%      0.47%        49     up_write                                anon_vma_prepare (20)
     0.44%      0.44%        45     free_pages_prepare                      free_hot_cold_page (1)
     0.50%      0.44%        45     get_pfnblock_flags_mask                 free_hot_cold_page (1)
     0.31%      0.44%        45     free_hot_cold_page                      free_hot_cold_page_list (4)
     0.85%      0.43%        44     next_zones_zonelist                     __alloc_pages_nodemask (20)
     0.24%      0.43%        44     kmem_cache_free                         putname (20)
     0.36%      0.42%        43     strnlen_user                            copy_strings.isra.22 (2)
     0.80%      0.42%        43     mem_cgroup_page_lruvec                  pagevec_lru_move_fn (3)
     0.89%      0.42%        43     mem_cgroup_update_lru_size              __pagevec_lru_add_fn (3)
     0.90%      0.40%        41     __slab_free                             kmem_cache_free (1)
     1.32%      0.40%        41     cmpxchg_double_slab.isra.54             __slab_free (1)
     0.30%      0.36%        37     __dec_zone_state                        __dec_zone_page_state (1)
     0.11%      0.36%        37     __dec_zone_page_state                   page_remove_rmap (1)
     0.18%      0.33%        34     vmacache_find                           find_vma (1)
     0.15%      0.33%        34     find_vma                                find_extend_vma (10)
     0.27%      0.29%        30     vma_compute_subtree_gap                 vma_adjust (10)
     0.04%      0.28%        29     flush_tlb_batched_pending               unmap_page_range (5)
     0.15%      0.27%        28     up_read                                 __do_page_fault (4)
     0.12%      0.25%        26     _find_next_bit.part.0                   find_next_bit (3)
     0.63%      0.25%        26     __rb_insert_augmented                   anon_vma_interval_tree_insert (15)
     0.11%      0.25%        26     cciva                                   main (1)
     0.31%      0.23%        24     mem_cgroup_from_task                    handle_mm_fault (1)
     0.13%      0.23%        24     prepare_exit_to_usermode                retint_user (1)
     0.16%      0.23%        24     find_next_bit                           cpumask_next_and (2)
     0.22%      0.23%        24     handle_mm_fault                         __get_user_pages (2)
     0.05%      0.22%        23     do_page_fault                           page_fault (1)
     0.16%      0.22%        23     __do_page_fault                         do_page_fault (1)
     0.76%      0.22%        23     down_read_trylock                       __do_page_fault (1)
     0.26%      0.22%        23     strcmp                                  find_next_iomem_res (1)
     0.05%      0.22%        23     error_entry                             page_fault (1)
```

Fig. 5 Extended simple-PT execution screen

3 Extended Flow Tracer Using Simple-PT

3.1 Data Structure

As shown in Fig. 4, we have proposed a basic data structure for making flow tracer based on PT dump information [8].

Figure 5 shows an example of running a program that extends simple-pt. As shown in Fig. 5, the extension program displays in real time the name and number of functions executed for a certain period, the relationship and number of calls between caller-callee, and the time occupied by the corresponding command in the entire command. The Simple-pt extension program analyzes the results of the simple-pt sptdecode program in real-time to produce a function call traceable form. Figure 6 shows a portion of the source of the simple-PT extension program.

```
    /* Extract INSNs */
    j = k = f_location = 0;
    for (i = 0; i < strlen(string); i++){
        if(string[i] == '[') j = i;
        if(string[i] == ']') k = i;
        if(string[i] == '>') f_location = i;
    }
    if(f_location == 0) continue;
    q = (INSN_WIDTH -1);
    for(i=0; i< INSN_WIDTH; i++) a_insn[i] = '0';
    for(i = (k-1); i > j+1; i--){ a_insn[q--] = (string[i] == ' ') ? '0' : string[i]; }
    cur_insn = atoi(a_insn);
    /* caller extraction */
    j = skip = 0;
    for (i = k+1; i < (f_location -2); i++){
        while (i < strlen(string) && string[i] != ' ' && string[i] != '+' &&
                (isalnum(string[i]) || string[i] == '_' || string[i] == '.')) {
                caller[j++] = string[i++];
        }
        if(j != 0){
                caller[j] = '\0';
                break;
        }
    }
    if(isdigit(caller[0])) continue;
    j = 0;
    if(isdigit(string[f_location+1])) continue;
    if(f_location == 0)     f_location = strlen(string) - 1;
    for (i = f_location+1; i < strlen(string); i++){
        while (i < strlen(string) && string[i] != ' ' &&
                (isalnum(string[i]) || string[i] == '_' || string[i] == '.')) {
                unit[j++] = string[i++];
        }
    if((j != 0) && (! isdigit(unit[0]))){
        unit[j] = '\0';
        total_symbol++;
        count = update(s, unit, caller, cur_insn, count);
        j = 0;
    }
}
/* Sort the structure in an descending order */
for(i=0; i<count; ++i){
    for(j=(i+1); j<count; ++j){
        if(s[i].insn < s[j].insn){
                temp = s[i];
                s[i] = s[j];
                s[j] = temp;
        }
    }
}
```

Fig. 6 Extended simple-PT decoder sample code

4 Conclusion

Hardware tracing is an important tool. Hardware tracers include ARM Coresight and Intel PT. A hardware tracer generates its enormous data into a log that is used for both performance analysis and debugging. PT is a new hardware-based tracing feature for Intel CPUs that traces branches executing on the CPU, which allows the reconstruction of the control flow of all executed code with minimal labor. PT can lookout the instructions of the user and kernel-level code, and can collect cycle information up to the detailed steps of instruction branches. simple-pt is a simple implementation of Intel PT on Linux. simple-pt decodes the branch trace and displays a function or instruction level trace. In this paper, we implemented an in-line tracer that extends simple-pt in Linux. The in-line tracer proposed in this paper provides real-time information of symbol, specific address occurrence, frequency, overhead, and caller/callee based on information provided by simple-pt.

Acknowledgements Funding for this paper was provided by Namseoul university.

References

1. Suchakra S (2015) Hardware tracing for fast and precise performance analysis. https://thenewstack.io/hardware-tracing-fast-precise-performance-analysis/
2. https://sites.google.com/site/intelptmicrotutorial/
3. Andi K, Adding processor trace support to linux. https://lwn.net/Articles/648154/
4. Andi K, Simple intel CPU processor tracing on Linux. https://github.com/andikleen/simple-pt
5. Peter T, Intel processor trace: how to use it (2016). https://tthtlc.wordpress.com/2016/01/26/intel-processor-trace-how-to-use-it/
6. Andi K simple-PT. https://github.com/andikleen/simple-pt
7. Thalheim J, Bhatotia P, Fetzer C (2016) INSPECTOR: data provenance using intel processor trace (PT). International conference on distributed computing systems (ICDCS), pp 25–34
8. Kim H, Kim Y, Kim I, Kim H (2017) Dynamic information extraction and integrity verification scheme for cloud security. In: Lecture notes in electrical engineering, vol 425. Springer

The Effect of Wearing the Customized Insole on the Coordination of the Right and Left Wrists Measured by Wearables During Golf Swing

Kyungock Yi and Hyeonjung Oh

Abstract The purpose of this study was to investigate the effect of the presence or absence of the insole on the intra—limb coordination between right and left wrists measured by IMU wearables. The subjects were one male and one male with foot deformity. The IMU-based wearables (CUBE MOTION, K2155~58, MSIP-CRM-sed-CM-IMUI-T) sampling rate was 800 Hz/s. The customized insole was constructed to raise the inner arch so that the RCSP would be in neutral position. The custom foot corrective insole influenced the proximal wrist movement during the golf swing. It was also effective in correcting the golf performance and as a result increasing the distance and reducing the angle of flight. Therefore, the acceleration-based IMU can conclude that it is a reliable and valid tool for kinematic analysis. This study will contribute to provide a basis for evaluating the validity of kinematic analysis using IMU wearables as well as identifying the effects of customized insole on golf performance.

Keywords Insole · Right and left wrist · IMU wearables · Coordination

1 Introduction

The three elements of golf are power, accuracy and consistency. These three elements are only possible with a strong foundation to use and adjust the force. The basis of this is ground reaction force. The ground reaction force is the starting point for creating the setup and swing [1]. Furthermore, ground is a fundamental element in developing and maintaining a swing. Therefore, it is no exaggeration to say that these three factors depend on the ground reaction.

K. Yi (✉) · H. Oh
Division of Kinesiology and Sports Studies, College of Science
and Industry Convergence, Ewha Womans University, Seoul, Korea
e-mail: yikok@ewha.ac.kr

H. Oh
e-mail: ohjung0723@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_10

In the case of beginners, the ground reaction force cannot be used properly and stably. That is, it moves without fixing the foot on the ground. In particular, golf players need to adapt to the various materials and grades of the ground, so training of the proprioceptive system of the foot is necessary.

Unfortunately the foot of a golf player has a lot of flatness because of the force that gives strength to the inside of the knee when swinging. There is a foot deforming mechanism of a golf player. During the swing, the knee is positioned more inward than the ankle, so the downward force on the swing is applied to the inside of the foot. For this reason, training methods have been devised to prevent the knees from gathering inward enough to move away from the center axis during the backswing. Wedge shoot and putting are not as wide as stride, so you can align your knee and ankle with this training method. However, the iron and driver shots have a wide stride and a strong power, and due to the nature of the swing that needs to be rotated, the vertical line from the knee downwards into the ankle. With this mechanism, the inner arch of the foot collapses, which increases the possibility of becoming flat.

Foot deformations of a golf player can lead to errors in the proprioceptive system of the foot, which cannot accurately receive information from the ground and, as a result, can lead to errors in the body's neuromuscular control [2].These errors can cause discomfort, pain and injury [3]. Customized corrective insole can make golfer help to rebuild the reactive neuromuscular control by using the proprioceptive feedback information from the foot [4, 5].

Golfers who has a pes planus can improve their errors and gain a more efficient working pattern through a reactive neuromuscular control via customized corrective insole. Usually, the foot correcting insole helps to align the ankle and inner arch and corrects the movement of the rear foot to fore foot [6]. Furthermore, it helps to align the knee and pelvis. In addition, insole is recommended for people with diabetic foot as well as those with deformed feet. In particular, customized insole for diabetic patients is good for prevention of ulceration, reduction of peak pressure and patient-based response, and cost effectiveness.

The club face is dependent on the wrist, the wrist on the arm, the arm on the torso, the torso on the leg, the leg on the foot and the foot on the ground [1]. The ground does not move. If we reverse this approach, we can infer that the movement of the foot movement affects the legs, trunk, arms, wrist, and also affects the club face. But we can ask these questions. Will it affect the movement of people's wrist when people wear their insole on their shoes? If so, what mechanism can be said to affect it?

Golf is a fast action that takes less than 1.5 s, so there is a limit to the difficulty of identifying the intra limb coordination between right and left wrist through cinematographic analysis. Because the wrists overlap each other during golf swing. So far, research on the wrist in golf swing has been limited to intra-segmental coordination analysis of the left arm. The movement of the left wrist is important to keep the left wrist flat during swing (Flat Left Wrist). This means that the club shaft and the left arm are aligned in a three-dimensional space, and the golf swing is a condition for making the left arm, club shaft, and club face all in-line on the coronal and sagittal planes. Therefore, the motion of the left wrist is an important factor in determining the direction of the ball. The right wrist moves in accordance with the

Table 1 Characteristics of the subjects

	Sex	Age	Height (cm)	Weight (kg)	Career (years)	Handy	Case history	RCSP (°)		Forefoot to rearfoot	
								L	R	L	R
sub 1	Female	38	172	67	23	9	5th, 6th on lumbar disc herniation	−5	−3	−4	−4.5
sub 2	Male	46	178	98	21	5	R. ruptured the posterior cruciate ligament	−4	−2	+4	+2

lead of the left wrist and serves as a support. The right wrist also plays an important role in delivering power from the body to the golf clubs. The right wrist has dorsi flexion and radial deviation in the backswing. However, it is necessary to identify the three-dimensionally coordinated relationship between the two wrists to determine the successful swing.

The purpose of this study was to investigate the effect of the presence or absence of the insole on the intra—limb coordination between right and left wrists measured by IMU wearables. This study will contribute to provide a basis for evaluating the validity of kinematic analysis using IMU wearables as well as identifying the effects of customized insole on golf performance.

2 Method

Subject: There were two subjects. One was a 38-year-old female, 67 kg, 172 cm, and had a golf career of 23 years. She had her fifth and sixth lumbar disc herniation. Her handy was nine. Her RCSP was left −5, right −3. Her forefoot to rear foot was left −4, right −4.5. The other was male, 46 years old and 75 kg, 178 cm who had 21 years golf experience, diabetes, back pain, redness edema foot, a ketone flavor, and forefoot varus. He ruptured the posterior cruciate ligament on the right knee in a traffic accident 11 years ago. His handy was five. His RCSP was left −4, right −2. His forefoot to rear foot was left +4, right +2.

The insole was adjusted to a shore. A hardness of 30–45° in consideration of the weight of the subject (Tables 1, 2, 3 and 4).

Research Question:

- Will wearables based on acceleration possible to perform kinematic analysis of both wrists according to the phases during golf swing?
- Does the presence or absence of an insole affect the coordination of both wrists?
- Does the difference in coordination between the two wrists affect distance, speed, and angle depending on the presence or absence of the insole?

Table 2 Difference according to kind of insole

	Corrective insole	Sports insole	Developed insole in this study
Cell	Polypropylene (4 mm)	Polypropylene (3 mm)	Polypropylene (4 mm) shore A 45~5
Posting	EVA	EVA	EVA shore A 30–35
Cover	Prescription according to disease	Lots of cushioning for shock absorption	Multi-form (3 mm) shore A 15–20, Does not give a lot of cushion feeling
Angular prescription principles	• RCSP neutral • Pain control strategy	• Total contact strategy • Reduced muscle fatigue	• RCSP neutral • Motion control and motion enhancement
Effect	• Reduced pain • Posture correction (growth period)	Reduced partial fatigue and pain	• Posture correction • Motion control and enhancement • Reduced fatigue (Indirect purpose)
Disadvantages	Hardness/difficulty in fitting shoes	Increased fatigue weak ground reaction force and weakness of sports technology improvement	Sophisticated adjustment failures may limit sporting behavior causes for pain (need adjustment)

Table 3 Results of subject 1 with or without insole

	Distance (m)	Ball speed (m/s)	Club speed (m/s)	Trajectory angle (°)
Without insole	117.44	39.70	32.50	27.30
	∧	‖	∧	∨
With insole	119.46	39.70	32.80	24.20

Table 4 Results of subject 2 with or without insole

	Distance (m)	Ball speed (m/s)	Club speed (m/s)	Trajectory angle (°)
Without insole	157.85	48.2	37.1	20.4
	∧	∧	∧	∨
With insole	190.61	56.30	41.40	18.20

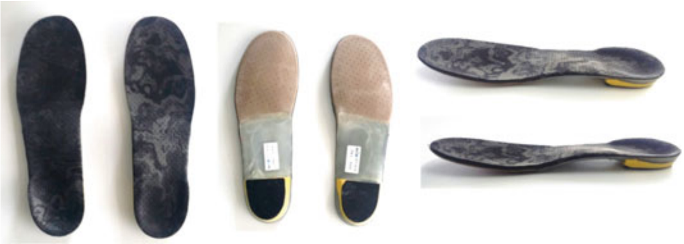


Fig. 1 The front, back, and side of the developed insole

3 Procedure

High speed camera was used with 1000 Hz/s of filming rate. The IMU-based wearables (CUBE MOTION, K2155~58, MSIP-CRM-sed-CM-IMUI-T) sampling rate was 800 Hz/s. The data were smoothed to 150 Hz.

Foot deformity was measured using Resting Calaneal Stance Position (RCSP) and forefoot to rearfoot. RCSP is the angle between the bisector of the calcaneus and the ground after the line is drawn. −means inside, and + means outside. +, −2° or more were seen as requiring correction, and both subjects were flattened, especially to the left by −5°, −4° respectively.

A custom insole for golf shoes was made after plaster casting of the subject's feet. The customized insole was constructed to raise the inner arch so that the RCSP would be in neutral position. Cells that serve as the skeleton of the middle and posterior regions was of a hard material (Polypropylene, 4 mm), and posting of the posterior was of a material that was softer (EVA) than the cell. The foot covering was made of soft multi-foam. The existing sports insole is made of a soft material and provides a cushioning feeling for shock absorption. Such insoles increase the fatigue of the foot as the game progresses because the insole is made to have a larger supporting area. Moreover, since it is a ductile material, it cannot accurately provide the ground reaction force and it is difficult to improve the golf skill. The customized insole of this study uses rigid insole for corrective purposes and improves the motion control and performance by utilizing accurate ground reaction force as well as immediate posture correction. The material, angle, effect and disadvantages of the foot insole, the existing sports insole, and the insole made in this experiment were summarized (Figs. 1, 2, 3, 4, 5, 6, 7 and 8).

Acceleration-based IMU devices were used and the sampling rate was set at 800 Hz. Two sensors were worn using a wrist strap. A reference point was set at the address posture. The acceleration in the right direction of the X axis is set to +, and the acceleration in the left direction is set to −. The acceleration in the front of the Y axis is +, and the acceleration in the rear is −. The upward acceleration of the Z axis is +, and the downward acceleration is −. The subjects were allowed to attach IMU devices and swing 10 times, depending on whether they were wearing custom

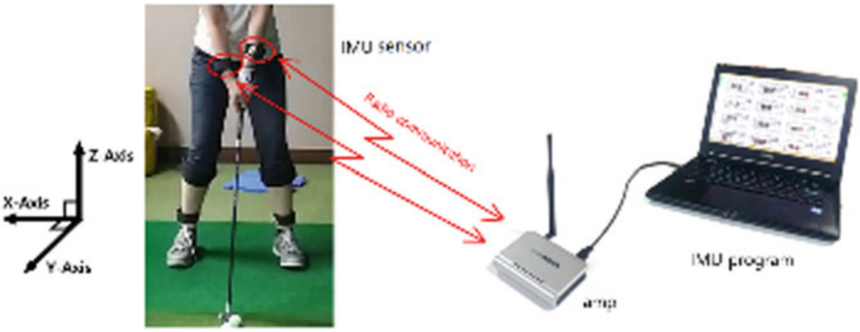


Fig. 2 Golf swing analysis system

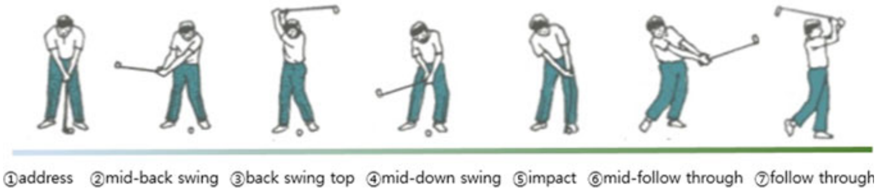


Fig. 3 Events of the golf swing

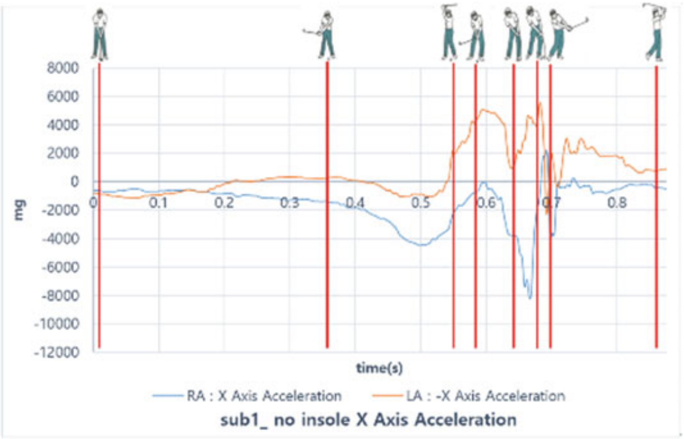


Fig. 4 X axis acceleration, without wearing the insole

insole after having practiced enough. The swing was evaluated according to the feel of the subjects hit, and the successful results were selected and analyzed. The event was divided into 7 events (Figs. 9, 10, 11, 12, 13, 14 and 15).

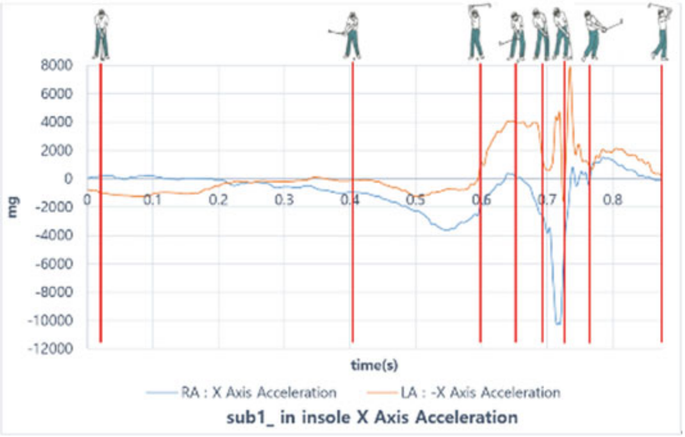


Fig. 5 X axis acceleration, with wearing the insole

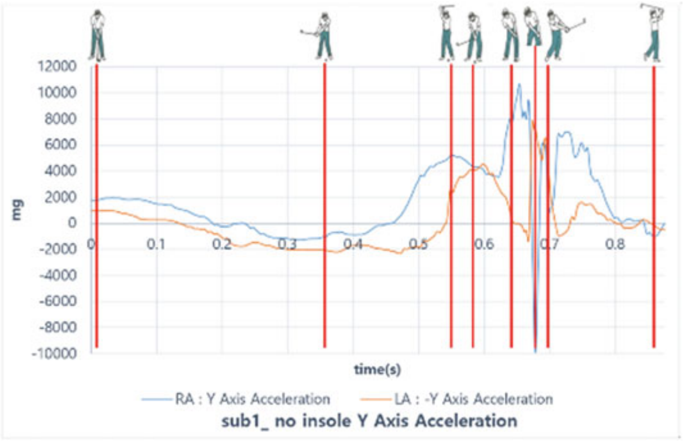


Fig. 6 Y axis acceleration, without wearing the insole

4 Result

1. The first subjects had increased distance and club speed after wearing insole, while the trajectory angle decreased. This means that the ball does not come up and the ball is farther away at a higher speed.

Regardless of the presence or absence of the insole, the coordination structure of the right and left wrist in the x direction exhibits the same relationship until the impact, while the acceleration of the left wrist was inversely related to the impact during the impact. However, the difference in the presence or absence of the use of the insole was during the mid-follow through. When the insole was worn, it was

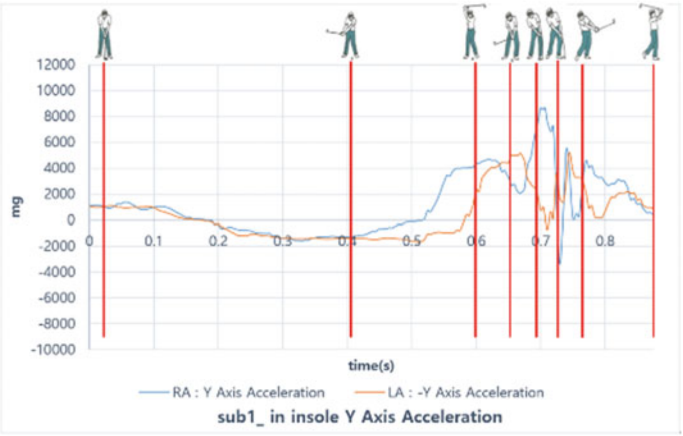


Fig. 7 Y axis acceleration, with wearing the insole

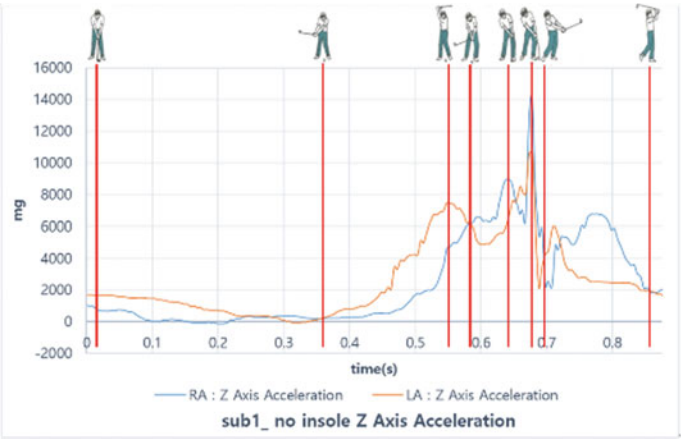


Fig. 8 Z axis acceleration, without wearing the insole

restored to the same relationship again during the mid-follow through. However, when the insole was not worn, there was an inverse relation between the right and left hands.

The acceleration graph in the X-axis direction shows a plateau phenomenon in the downswing at the left wrist when the insole was worn. This means that when the insole was worn, it moves consistently and stably without changing the speed.

In the Y-axis acceleration graph, a stable state, that was, a plateau phenomenon, occurred in the backswing and the downswing in the right wrist when the insoles were worn.

In the Z-axis acceleration graph, when the insole was worn, a high peak appeared at the moment of supporting the left leg, which was the last moment of the immediate

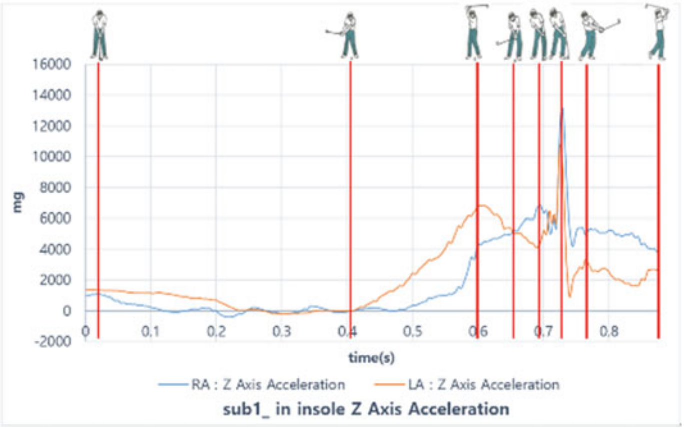


Fig. 9 Z axis acceleration, with wearing the insole

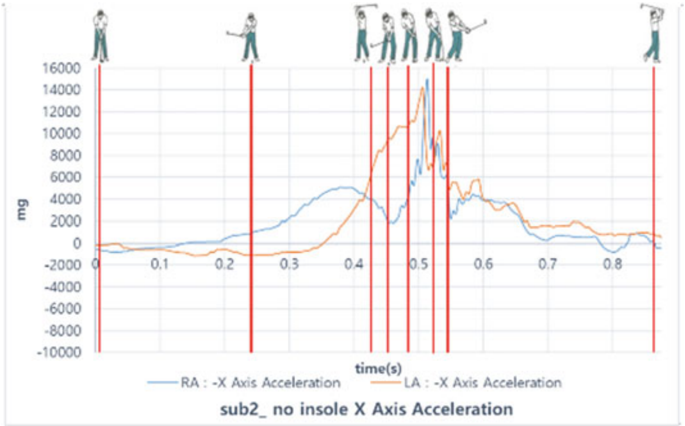


Fig. 10 X axis acceleration, without wearing the insole (subject 2)

impact zone. This was the moment when the ball and club were pushed together in the tangential direction of the swing orbit.

Immediately after impact with the mid-back swing, you will see a total of three right and left wrist crossings on the mid-follow through when not wearing the insole. It means turning the wrist, i.e. flipping. However, after wearing the insole, flipping only occurred in the mid-back swing. Flipping during golf swing produces inconsistent contact and leads to fat shots, thin shots and high, weak hits.

It also brings a lose club head speed and control of the clubface [7].

When the insoles were worn, from the mid-down swing to the mid-follow through, the left and right wrist move to the same relationship and the maximum acceleration of the left wrist increases. This can cause increased distance and club speed.

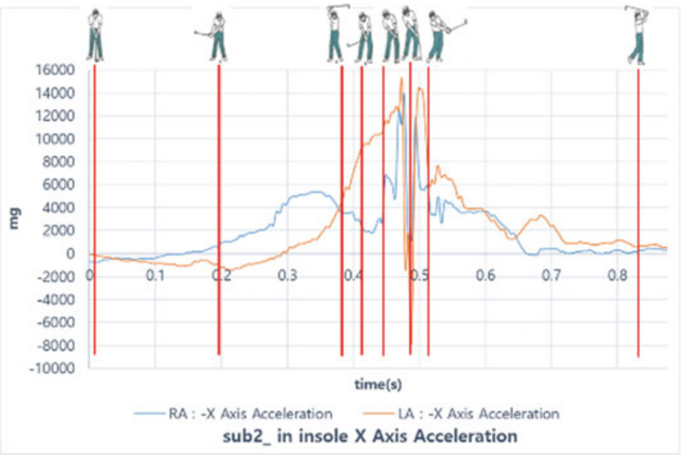


Fig. 11 X axis acceleration, with wearing the insole

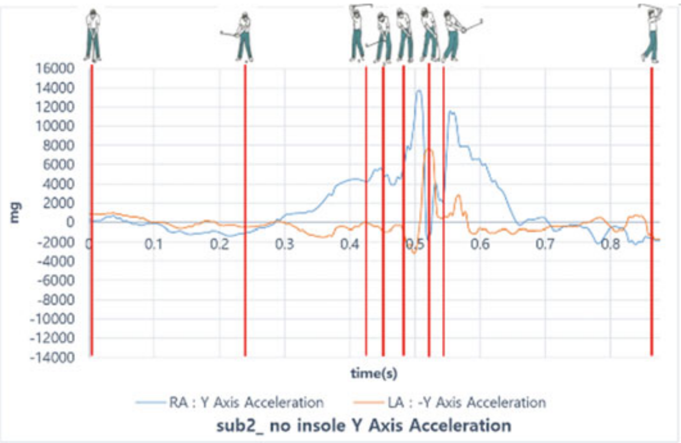


Fig. 12 Y axis acceleration, without wearing the insole

2. The second subjects also increased distance, ball speed and club speed when the insole was worn, while the trajectory angle decreased.

The second subject had no left wrist acceleration change from the top of the backswing to impact. This means that the same acceleration was maintained while keeping the left wrist flat. When the insole was worn, the second subject's swing pattern was characterized by the acceleration in the x, y, and z directions of the left wrist in the immediate impact zone [8].

The acceleration in the X direction of the left wrist was increased when the insole was worn, between the impact and the left leg support, and between the left leg support and mid- follow through in the immediate impact zone. The acceleration in the Y

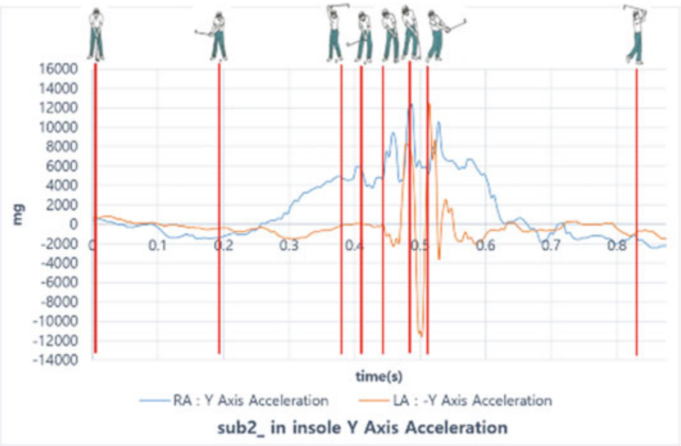


Fig. 13 Y axis acceleration, with wearing the insole

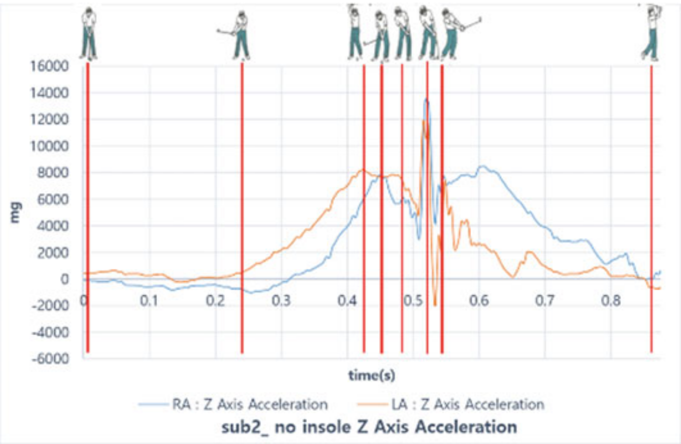


Fig. 14 Z axis acceleration, without wearing the insole

direction of the left wrist increased at the mid—follow through. The acceleration of the left wrist in the Z direction was higher than that of the right hand at the last moment of the immediate impact zone when the insole was worn, and the acceleration of the left wrist was the maximum at this time. This was the moment when the left wrist was released to maximize the centrifugal force in the tangential direction of the golf swing orbit, and was the left leg support phase in the immediate impact zone. This was the optimum hand release action. The primary purpose of the optimum hand release action is to control the clubface so that a golfer can hit the ball straight, and that means that any optimum hand release action must ensure that the FLW/clubface both face the target during their passage through the immediate impact zone [8].

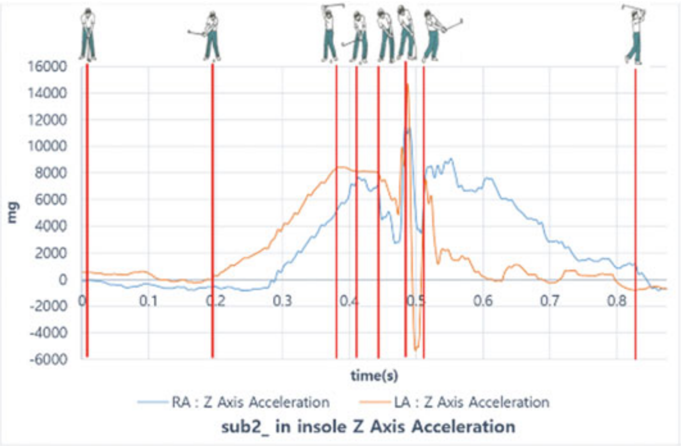


Fig. 15 Z axis acceleration, with wearing the insole

5 Result and Suggestion

The custom foot corrective insole influenced the proximal wrist movement during the golf swing. It was also effective in correcting the golf performance and as a result increasing the distance and reducing the angle of flight. Therefore, the acceleration-based IMU can conclude that it is a reliable and valid tool for kinematic analysis.

In the kinematic analysis of golf, in the immediate impact zone, the left leg support phase, needs to be added.

Future research on the effects of customized insole, upper limb movements, and long-term effects of insole wear is a challenge.

References

1. Bradly N (2004) Seven Laws of the Golf Swing, BBC Worldwide Limited
2. Brantingham JW, Chang MN, Gendreau, Price JL (2007) Case report, “The effect of chiropractic adjusting, exercises and modalities on a 32-year-old professional male golfer with hallux rigidus”. Clin Chiropr 10(2):91–96
3. Kroemer KHE (1989) Cumulative trauma disorders: their recognition and ergonomics measures to avoid them. ELSEVIER Appl Ergon 20(4):274–280
4. Voight ML, Cook G (1996) Clinical application of closed kinetic chain exercise. J Sport Rehabil 5(1):25–44
5. Cook G, Burton L, Fields K (1999) Reactive neuromuscular training for the anterior cruciate ligament-deficient knee: a case report. J Athl Train. 34(2):194–201
6. Chang C-J, Yang S-W, Chang C-W, Lin Y-L, Kuo F-C, Lin C-C, Liu K-T (2015) Effect of functional foot orthotics on golf swing stability and accuracy of shots. Footwear Sci 7(1):157–159
7. Raudenbush D (2017) One-plane golf swing fundamentals. Last Updated 11 Sept 2017
8. Mann J (2018) Perfect golf swing review: a critical review of the golf swing. http://perfectgolfswingreview.net/index.html. Accessed 19 Feb 2018

A Study on Development of Application for Management of Obesity in Children with Intellectual Disabilities

SeungAe Kang

Abstract The purpose of this study was to provide information necessary for development of the applications that could facilitate the management of obesity in children with intellectual disabilities and to present improvement measure for development of practical applications. The five contents of the application for management of obesity in children with intellectual disabilities are as follows: Promotion of intake of healthy foods, Promotion of physical activity, Early childhood diet and physical activity, Health/nutrition and physical activity for school-age children, and Weight management. In addition, this study presented the user expansion and their continuous involvement through interface simplification, alert function, reward, voice recognition and subtitle support strategies in consideration of the characteristics of the children with intellectual disabilities.

Keywords Obesity · Interface · Application · Feedback · Children with intellectual disabilities

1 Introduction

The use of mobile media has become part of daily life in this era of smart devices and technologies. Mobile media has taken a firm root in our lives to an extent that people sometimes experience new symptoms, such as 'Nomophobia Syndrome' which is a kind anxiety caused by the fear of unavailability of or inaccessibility to mobile phone. The difference between smartphones and traditional communication tools lies in the fact that smartphones can be used concurrently with a variety of applications to achieve specific purposes [1]. Mobile application refers to application software used directly by users on smartphone or tablet PC platforms [2]. In Korea, various

Funding for this paper was by Namseoul University.

S. Kang (✉)
Department of Sport and Healthcare, Namseoul University, Cheonan, South Korea
e-mail: sahome@nsu.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_11

applications have been developed and distributed rapidly since the first application was unveiled in the latter part of 2009, and furthermore, the proportion of smartphone users in Korea has expanded fast to exceed 79% of total population as of 2012 [3]. As the applications are free from time and space constraints, they are easily accessible to users and allow individual users to achieve their goals across various fields with minimal input of time and efforts, thus bringing maximum effect at minimum cost.

Mobile technology that allows individuals to select and manage the time and location conveniently is recognized as an effective approach to obesity management [4–7]. The "Noom Diet Coach", developed by a domestic IT company, is one of the most typical obesity management applications that help more than 19 million users around the world in the prevention of the risk of various chronic diseases caused by obesity as well as dieting and to lead healthy lives. Obesity is known to cause metabolic diseases, including hypertension, diabetes, and hyperlipemia, due to a decrease in physical function such as weight gain, increase in cholesterol level, decrease in physical activity, and decline in physical strength, etc. [8], which will exert a considerable socio-economic burden in the future. As the obesity in children and adolescents is highly likely to be developed into obesity in adulthood, efforts should be made to prevent and control childhood obesity.

Intellectual disability is defined as a disorder in which the IQ is less than 70 and the disorder is definitely determined before the age of 18 and poses significant difficulty in terms of cognitive function and adaptive behavior [9]. Children with intellectual disabilities tend to exhibit low self-protection ability, low communication skill, and low social adjustment ability, although those abilities are necessary for independent life, due to their behavioral changes and decline in their ability to perform roles [10], and furthermore, have relatively higher rate of obesity as a result of inadequate physical and social activities, compared with children with other types of disabilities [11–15]. Obesity in children needs to be managed systematically, considering that the obesity in children with intellectual disabilities has negative effects on health fitness, physical development, cognitive ability, and perceptual motor development [8], and increases the likelihood of adult obesity and chronic diseases.

Recently, various applications that target the children with intellectual disabilities have been developed in special education and athletic fields. Moreover, researches have been conducted on applications that tap into the social network or incorporate the augmented reality programs. The use of mobile technology as part of intervention approach for the management of obesity in children with intellectual disabilities is considered to represent a novel way to mitigate and resolve the problems of existing practices of obesity. The purpose of this study was to provide information necessary for development of the applications that could facilitate the management of obesity in children with intellectual disabilities and to present improvement measure for development of practical applications.

2 The Model for Approach to Obesity Management Using the Applications

Although various approaches have been attempted to the management of obesity in children with intellectual disabilities, there has been a growing demand for development of new approaches from the aspects of time/space constraints and cost effectiveness. With widespread distribution of smart phones, electronic communication technology has emerged as a new approach that can be leveraged to prevent and manage the obesity in children with intellectual disabilities. High smartphone penetration rate in Korea and burgeoning interest in smartphones among children with disabilities have created an environment conducive for multi-faceted and comprehensive approaches.

Application-based management of obesity has the following advantages [16]: Application-based management of obesity enables cost effectiveness, real-time data collection and feedback, reduction of participants' time-related burden and financial costs, flexible operation of programs, application to various environments/targeted people in many different age groups, and quick dissemination and distribution, compared with existing face-to-face approach. However, it should be also considered that there are also downsides to this novel form of intervention. The greatest downsides include the excessive use, reckless exposure to the media [17] which arise from easy access, and inappropriate content application [18]. Guidance would need to be provided in such a way that positive aspects can be highlighted based on selection of suitable contents and adjustment of the time for utilization.

As the obesity is attributable to the imbalance between energy intake and consumption which arises in connection with social and environmental factor, as well as individual factor (Fig. 1), comprehensive systematic approach would be needed, such as the increase in nutrition and physical activities (Fig. 2).

3 Components and Strategies of Applications

The obesity management applications, developed previously, consist of nondisabled user-oriented interfaces, making it somewhat difficult for children with intellectual disabilities to use them. Thus, it would be necessary to build efficient content elements suited for the characteristics of children with intellectual disabilities.

- Simplification of interface: Simple and intuitive interfaces are required, rather than complex contents, for children with intellectual disabilities who have a lack in verbal skill and poor attention which erode their learning ability and complicate their self-initiated learning activities. That stems from the fact that the interfaces of existing applications targeted for non-obese children do not induce the children with intellectual disabilities to think and make decision on their own in the performance of their roles. Thus, it would be necessary to design simple interface and increase the ease of use based on simplified content arrangement which allows the

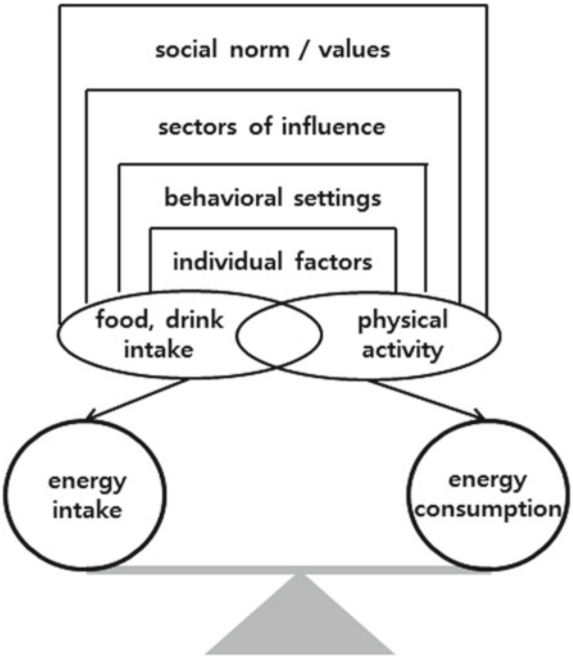


Fig. 1 CDC division of nutrition, physical activity and obesity, state nutrition, physical activity and obesity program (Technical assistance manual, January 2008) [16]

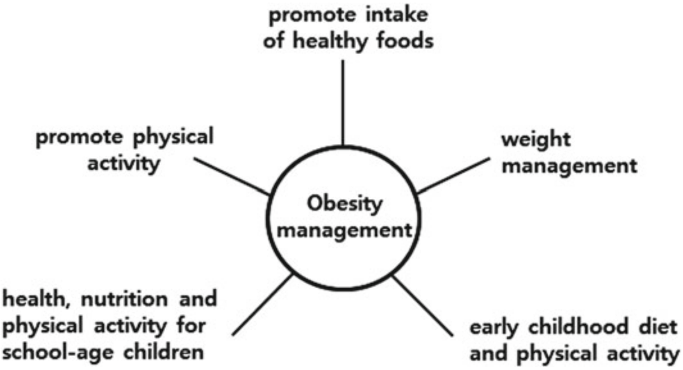

promote intake of healthy foods
promote physical activity
weight management
Obesity management
health, nutrition and physical activity for school-age children
early childhood diet and physical activity

Fig. 2 The five contents of the application for management of obesity in children with intellectual disabilities

children with intellectual disabilities to use independently without any problem by merely following the processes.

- Reminder function and reward: Behavior and attention of users can be induced by utilizing the alert functions that transmit to users the information such as the missions presented for obesity management while they are using the applications.

In addition, ranking may be provided based on fair competition with other users in daily missions where the scores are given through the gamification. In addition, users may be induced to increase participation and continue their involvement by allowing them to have more fun with avatar decoration, etc., on the basis of rewards given in the form of points or items whenever they achieve their missions or are ranked upward. Along with that, multimedia interaction elements may be introduced to create pleasure factors, like showing their changing avatars to each other, through the obesity management mission implementation.

- Speech recognition and subtitle support: The voice and subtitle functions of the contents for children with intellectual disabilities with low linguistic and voice expressive capabilities are likely to lead to a significant increase in users' utilization. Since those children face difficulty in understanding and using abstract language, voice and subtitle with relatively easy vocabulary will help increase the utilization. The voice recognition which activates application functions will also help children enhance their voice expression ability.

4 Conclusion

This study presented improvement measures for development of practical applications by providing strategic elements necessary for proper content configuration and development of applications designed to help manage the obesity in children with intellectual disabilities. It may be desirable to configure the applications for management of obesity in children with intellectual disabilities based on these 5 contents: Promotion of intake of healthy foods, Promotion of physical activity, Early childhood diet and physical activity, Health/nutrition and physical activity for school-age children, and Weight management. In addition, this study presented the user expansion and their continuous involvement through interface simplification, alert function, reward, voice recognition and subtitle support strategies in consideration of the characteristics of the children with intellectual disabilities. The obesity management applications to be developed in the period ahead will be able to be utilized as a novel approach to intervention which actively reflects the characteristics of children with intellectual disabilities.

References

1. Shin HJ, Lee HJ, Park JS, Jo HR, Na MJ, Cha SH, Kim DW, Park CW (2015) The investigational study on health-related mobile application software and its improvement. FDC Legal Res 10(1):1–9
2. Wasserman AI (2010) Software engineering issues for mobile application development. In: Proceedings of the FSE/SDP workshop on future of software engineering research, pp 397–400
3. Hong SP The smartphone distribution rate grew up to 80% this year. The Cellular News. http://www.cel\-lular.co.kr/37336

4. Kang JH (2014) Adolescent: obesity, internet/game addiction (Internet). Biotech Policy Research Center
5. Kim HK, Kim YS (2012) The effects of smartphone application to increase physical activity among university students. Korean J Phys Educ 51(5):457–466
6. Carter MC, Burley VJ, Nykjaer C, Cade JE (2013) Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. J Med Internet Res 15(4):32
7. Smith JJ, Morgan PJ, Plotnikoff RC, Dally KA, Salmon J, Okely AD et al (2014) Smart-phone obesity prevention trial for adolescent boys in low income communities: the ATLAS RCT. Pediatrics 134(3):723–731
8. Kim DM (2014) The effect of obesity rate on health-related physical fitness of people with intellectual disabilities. Korean J Adapt Phys Activity 22(3):15–28
9. Hong MY, Jung BK (2016) A systematic review on the effects of tablet PCs for children with intellectual disabilities. J Korean Soc Occup Therapy 24(3):67–79
10. Min CS (2003) The study on characteristic of mental retardation definition in historical change process. J Intellect Disabil 5:173–187
11. Min BI, Kim DC (2009) The effect of aerobic exercise on cardiovascular risk factors in mental retarded obese women. J Adapt Phys Activity 17(2):47–62
12. Park KY (2003) Pulmonary respiration-function and bone-density of mentally handicapped students, people with obesity and with no obesity. J Adapt Phys Activity 11(2):119–131
13. Son SH, Lee IK (2007) The effect of climbing as after-school activity on the body composition of obesity students with mental retardation. J Adapt Phys Activity 15(2):71–95
14. James H, Rimmer D, Fujura G (1993) Prevalence of obesity in adults with mental retardation. Am Assoc Mental Retard 31(2):105–110
15. Ponichtera J, Mathews T, Glaser R (1992) Maximal aerobic power of individuals with multiple sclerosis using ergometer exercise. Med Sci Sports Exerc 24:73
16. Kang JH (2014) BT-IT convergence based youth obesity prevention/management technology. Biotech Policy Research Center Specialist report, vol 2, pp 1–9
17. Yoo KJ (2010) A study on the development of program by using smart phones and tablet PC and its effects on scientific thinking of young children. J Korea Open Assoc Early Child Educ 6(17):85–110
18. Barr R, Danzinger C, Hilliare ME, Andolina C, Ruskis J (2010) Amount, content, and context of infant media exposure. Int J Early Years Educ 18(2):107–122

A New Paradigm for Spectator Sports in Application of Media

Minkyu Kim, Soojung Park and ByoungKwon Park

Abstract This study examines features of information technology that can provide new information for viewers of sportscasts as a means of spectator sports, whose demands are increasing as a leisure activity. Additionally, expected effects of such features are discussed. As information and communication technology advances, it is expected that checking psychological conditions in sports will be possible in the near future. This study designs a system that monitors players' psychological conditions and resulting changes in their motor functions by means of application technology. Such technology applications will be of significant values as new media contents as well as basic materials for players' performance enhancement.

Keywords Spectator sports · Application of media · Brain wave measurer
RFID chip

1 Introduction

Modern society is a leisure-centered society [1]. In the development of civilizations, time allotments, interests, and desires in human life have been rapidly changed. The advancement of medical science and science technology has increased leisure time significantly. Economic development, democratization, and globalization have

M. Kim · S. Park · B. Park (✉)
Graduate School of Education, Inha University, Incheon, South Korea
e-mail: zexrol@naver.com

M. Kim
e-mail: loisir@inha.ac.kr

S. Park
e-mail: psj@inha.ac.kr

M. Kim · S. Park · B. Park
Department of Kinesiology, Inha University, Incheon, South Korea

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_12

changed the former society where production led consumption into a society where consumption leads production and not labor but leisure is viewed as a major purpose of life [2].

Leisure has resulted in positive effects as the labor-centered, quantitative society became the leisure-centered society where quality and value are regarded as valuable in human life. Participation in leisure activity improves the quality of life as it contributes to physical health, social health in interpersonal relationships, educational improvement through new experience and satisfaction of curiosities, psychological health through self-realization and feeling of well-being, and so forth [3].

Among various types of leisure activities, the number of those enjoy spectator sports through media is increasing gradually, which indicates that many sportscast services other than the traditional broadcasting via TV are influential. Indeed, various forms of services are available: social cast by means of social media; text messaging services via mobile or Internet devices; and multi-angle broadcasting methods that provide images of sports games in many different angles [4]. As people recognized the positive effects of leisure and various related desires grew among them, the convenience of watching even on the way and fun factors of watching images in a variety of formats seem to satisfy viewers' demands.

Among various items of spectator sports, pro baseball is one of the most popular pro sports in Korea. According to KBO (2017), the number of spectators of baseball games as of 2017 is about 8.5 million. Many statistical materials also support this report [5]. At present, domestic pro baseball games are all broadcast live through various channels. Even highlight images and various related articles are far more than in other pro sports.

Information provided during domestic pro baseball broadcasts includes close-up videos of hitting moments, schedules, team ranks, detailed records, play information, etc. To assist referees' decisions, state-of-the-art 4D replay images and 'ultra-slow' images taken by means of ultra-high speed cameras of 400–1,200 frames are provided during broadcasts. Such new ways of broadcasting and reporting create more fun in sports broadcasts, making viewers more absorbed in spectator sports.

Applying such technologies to player training can improve training methods and players' performance. When it comes to players' performance in the area of sports science, exercise prescription for their physical activity and exercise has been recognized as the most important service. Exercise prescription means to determine the appropriate quality and amount of exercise depending on each participant's present physical stamina and provide programs of various exercise forms, strength levels, and periods. Such programs include health examination, physical strength test, Bruce protocol, etc. [6]. However, certain sports such as baseball require more than developing and applying programs such as exercise prescription that are designed merely to improve physical functions. Training methods focusing on improving running performance on a linear track cannot improve base running performance in baseball. Mental aspects emphasized in sports psychology cannot be addressed either. As such, various other determinants of performance are involved in sports game [7].

Accordingly, this study examines features of information technology that can provide new information for viewers of sportscasts as a means of spectator sports, whose demands are increasing as a leisure activity. Additionally, expected effects of such features are discussed.

2 Body

2.1 Technology of Collecting Brain Wave Data

Physical exercise analysis technology may utilize sensors that are attached on a human body in order to measure heart rates, blood pressure, respiration, calorie consumption, etc. through a wireless network. The participant is given information on his or her exercise strength, distance, quantity of motion, frequency, etc. Related devices for scientific sports activity have been commercialized.

Various information and communication technologies are utilized and applied in sports games, it is possible to monitor pro baseball players' conditions during a game real-time, which makes the game even more fun. In recent pro baseball broadcasts, for example, information on the pitcher's pitch patterns, catcher's positions, ball speed, etc. is utilized to present many interesting things to see. Furthermore, such information contributes to enhancing the game performance and scientific analysis of games with a higher chance of winning.

Physiological conditions of the body and psychological skills affect the players' performance during the game significantly. The psychological state during a game affects the player's technical motions, which are directly related to the program and energy generating system. Ultimately, the body, which is the basic execution system of technical motions is a deciding factor of motional efficiency [8]. Grasping the psychological state of players during a game is, therefore, a key element in a scientific approach to performance enhancement. It is important to check fundamental factors of players' psychological conditions so as to maximize their performance during a game. This is not only to improve the game performance but also make the game more exciting only if the spectators can see players' psychological conditions. To this end, brainwave technology and RFID technology may be applied to players in order to generate and process big data during games in combination with broadcasting technology.

In general, mental or psychological states of players in a stadium are likely to be varied and changing constantly [9]. While there are positive psychological elements such as confidence, morale boost, excitement, etc., negative elements of psychology such as anxiety, tension, anxiousness, and worry are more common [8].

Any player would go through such mental conditions during a game, and the types and intensity may change often over time [10] and depending on each player's personal characteristics.

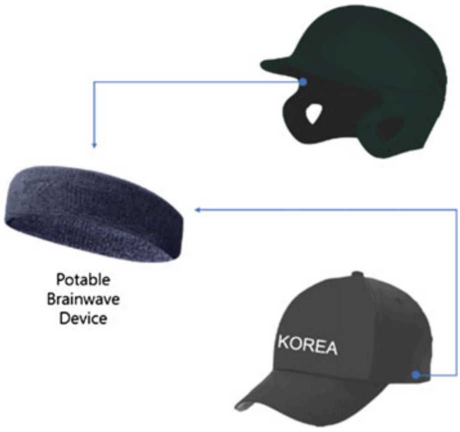


Fig. 1 Baseball helmet and cap in which a brain wave measurer is embedded

Players may experience drastic and sudden emotional changes during a game. For instance, psychological elements of a catcher or a pitcher when there are runners on some or all the bases would be different from those of a batter. In an important match of a pitcher and a batter that needs high concentration, those confident of one's motor ability would show a high level of Mid-β and γ waves, which indicate high concentration and stress respectively. As such, players' psychological conditions and interests are likely to be reflected in physical activity and can be measured. Figure 1 shows a batter's helmet and a pitcher's cap in which a brain wave measurer is embedded in order to check brain waves of players during a game.

2.2 RFID-Based Base Running Information Gathering

There are various areas where RFID (Radio-Frequency Identification) can be applied for identification. Major examples are transportation cards, credit cards, etc. which are used for payment, entrance, and certification.

RFID chips may be applied to baseball players' shoes, batters' gloves, and base readers which sense when a player arrives at the base. These chips make it possible to measure how fast players move, how long it takes before a pitcher throw a ball, and patterns a player tend to show at base running. As shown in Fig. 2, each player's personal ID is referred to and data is measured by means of RFID chips at the bottom of baseball shoes and batters' gloves.

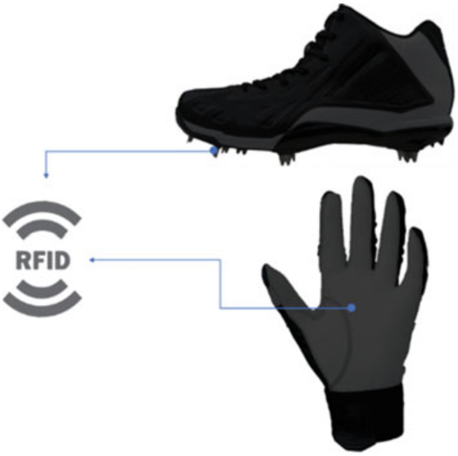


Fig. 2 Baseball shoes and gloves in which a RFID chip is embedded

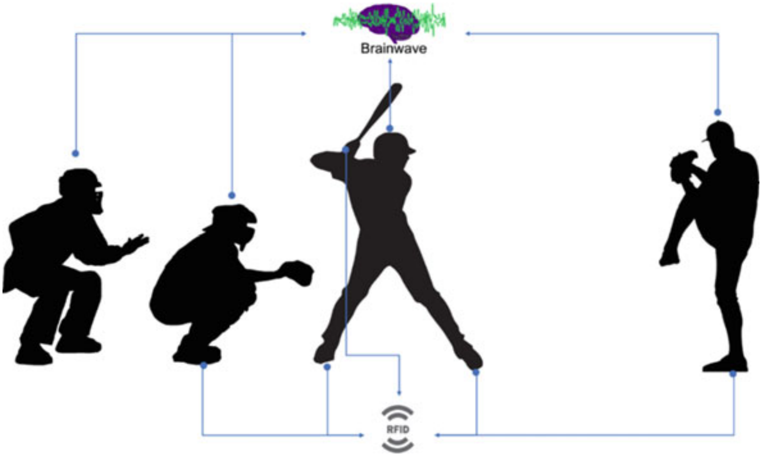


Fig. 3 The layout of information on game progress in application of brain wave measuring devices and RFID chips

2.3 How to Apply

This system helps grasping psychological states of baseball players and resulting changes in their movements. Figure 3 illustrates the basic layout that shows psychological conditions of the referee, catcher, pitcher, and player at bat. This will make the game more fun since spectators can see players' psychological states changing during the game depending on variables and predict who the psychological warfare between the pitcher and the batter would end.

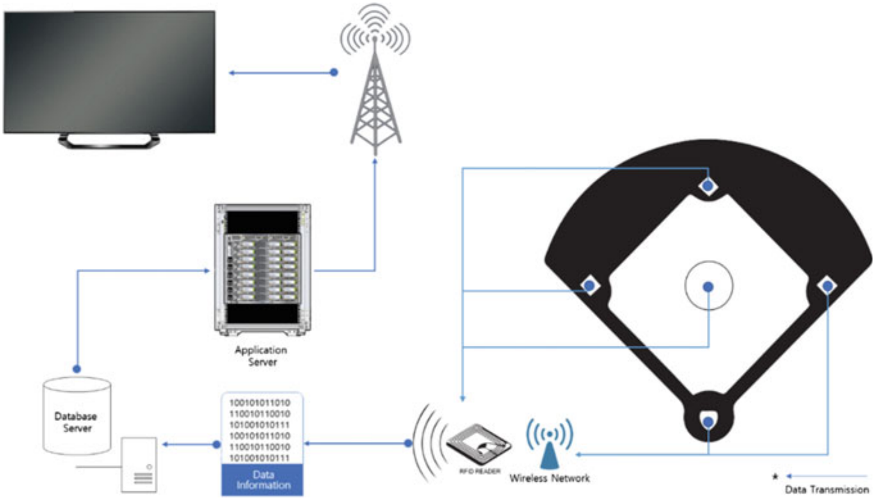


Fig. 4 The layout of the game data transmission system

As shown in Fig. 3, generated data is transmitted to the database server by means of RFID readers and the wireless communication network. The data is then transmitted to the application server for analysis, and the analyzed data is delivered to watchers through the network. Figure 4 shows flows of data, which can be utilized for spectators' fun and as a basis on which leaders can develop psychological training programs for their players.

3 Conclusion

This study examines features of information technology that can provide new information for viewers of sportscasts as a means of spectator sports, whose demands are increasing as a leisure activity. Additionally, expected effects of such features are discussed. To this end, the IT devices were utilized to collect and analyze data on emotional changes that baseball players would go through depending on the game phases.

Particularly, technologies to collect brain-wave and biological data and physical activity data were utilized in order to monitor various base-running aspects. Such technology applications will be of significant values as new media contents as well as basic materials for players' performance enhancement.

References

1. Kim M-K (2015) The relationships between serious leisure, recreation specialization and leisure addiction. J Leis Stud 13(1):89–104
2. Woo K-J (2009) Contemporary leisure & tourism. B&M books, Seoul
3. Kim M-K, Park S-J (2014) Grounded theoretical analysis on the formation of leisure addiction. J Leis Recreat Stud 38(3):1–16
4. Seo D-M, Kim S-H, Park H-G, Ko H-D (2012) Real-time text scoreboard system using social media and live media. Korea Inf Sci Soc 193–195
5. KBO home page (2017) Number of spectators by year. http://www.koreabaseball.com/History/Crowd/GraphYear.aspx
6. Kim M-K, Park S-J, Park B-K (2016) A new paradigm for the spread sport leisure culture focusing on the IT-based convergence interactive system. Lect Notes Electr Eng 376:1477–1485
7. Lee Y-H (2007) Empirical evidence on the determinants of team performance in Korean baseball league. J Korean Soc Meas Eval Phys Educ Sports Sci 63–77
8. Kim K-W (1999) Metal states of elite athletes in competition. Korean Alliance Health Phys Educ Recreat Dance 197–207
9. Anshel MH (1997) Sport psychology from theory to practice, 3rd edn. Gorsuch Scarisbrick Publishers, Scottsdale, AZ
10. Catty BJ (1983) Psychology in contemporary sport, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Physical Activity Intervention Platform for the Elderly with Mild Cognitive Impairment

SunYoung Kang and SeungAe Kang

Abstract The purpose of this study is to propose a physical activity platform as an intervention method to induce the increase of physical activities of the elderly with mild cognitive impairment, a pre-dementia stage leading to actual dementia which emerged as social problem amid the increase in the elderly population. Physical activity plays a role in improving cognitive function and also serves an important role in mitigating and preventing the decrease in cognitive dysfunction arising from mild cognitive impairment and aging, and it has been found that improvement in cognitive function is observed more frequently in older elderly than in relatively young elderly. Vigorous efforts need to be made for management of the elderly with mild cognitive impairment based on formation of efficient platform structure which helps increase physical activity, which is the main factor preventing the progression to dementia. The data on physical activity and health, collected from the sensors detecting the amount of exercise, mobile devices and wearable device outfitted with exercise sensors, are periodically updated through the platform to provide information to the elderly and their families in an open and bi-directional way. Moreover, the two-way information which is provided based on feedback process can help increase the physical activity, enabling prevention of decline in cognitive functions and promoting maintenance of cognitive functions.

Keywords Wearable device · Mobile · The elderly · Platform · Mild cognitive impairment · Physical activity

S. Kang
Department of Physical Education, Korea University, Seoul, South Korea
e-mail: 1010kang@hanmail.net

S. Kang (✉)
Department of Sport and Healthcare, Namseoul University, Cheonan, South Korea
e-mail: sahome@nsu.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_13

1 Introduction

As the aging begins, most functions, such as sensory, cognitive, and physical functions, starts to be impaired [1]. In particular, changes in sensory and cognitive functions are important factors affecting the use of the Internet by the elderly, who also face difficulties with the use of information due to the decline in their visual and cognitive abilities [2–4]. Nevertheless, the Internet utilization rate and the mobile device retention rate among the elderly over 65 years of age have been on an upward trend. 7 out of 10 people (74.5%) among those in their 60s were found to be internet users. Internet utilization rates among the elderly aged 70 or higher currently stand at 25.9%, increasing steadily from 14.1% in 2014 and 17.9% in 2015. The utilization rate of smartphones and wearable stands at 64.1% among those in their 60s and 14.9% among those in their 70s [5], which has been steadily rising. Active seniors who are physically healthy and active compared to their actual ages are familiar with IT smart devices such as smartphones and computers and tend to use them in everyday life as actively as young people [6].

The mild cognitive impairment (MCI) refers to a state in which cognitive function, particularly memory, is reduced compared to that of same age group while the ability to perform everyday life is preserved. In other words, early management is important to ensure early detection and treatment in order to maximize therapeutic effect, considering that the person with mild cognitive impairment has yet to reach the stage of dementia but remains in an intermediate stage between normal aging and dementia which can progress to dementia [7]. MCI requires continuous management through cognitive stimulating activities, physical exercises/activities, and dietary factors recommended as protective factors in addition to treatment such as medication and vascular risk factor control, etc. Exercise above medium intensity increases the Brain-Derived Neourotrophic Factor (BDNF) level and promotes creation of new neurons and maintenance of existing neurons. As active physical activity is an important factor for maintaining cognitive function, various approaches are needed to induce active participation in physical activities.

Smartphone has become a necessity for the elderly, too. In fact, the increasing number of adults aged 50 or older who use smartphones freely shows the possibility that the elderly-related services and health care paradigm can be changed in the future. Thus, in this study, we intended to propose a physical activity platform as an intervention method to induce the increase of physical activities of the elderly with mild cognitive impairment, a pre-dementia stage leading to actual dementia which emerged as social problem amid the increase in the elderly population.

2 Cognitive Impairment and Physical Activity

It is not easy to determine the borderline between cognitive dysfunction and mild cognitive impairment which occurs commonly in normal aging processes. We should suspect mild cognitive impairment if a person experiences a significant decline in

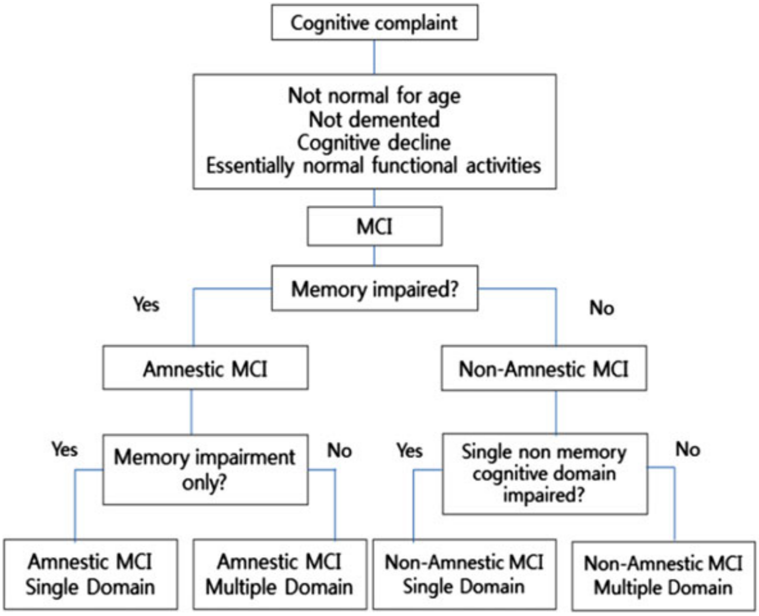

Cognitive complaint
Not normal for age
Not demented
Cognitive decline
Essentially normal functional activities
MCI
Memory impaired?
Yes
No
Amnestic MCI
Non-Amnestic MCI
Yes
Memory impairment only?
No
Yes
Single non memory cognitive domain impaired?
No
Amnestic MCI Single Domain
Amnestic MCI Multiple Domain
Non-Amnestic MCI Single Domain
Non-Amnestic MCI Multiple Domain

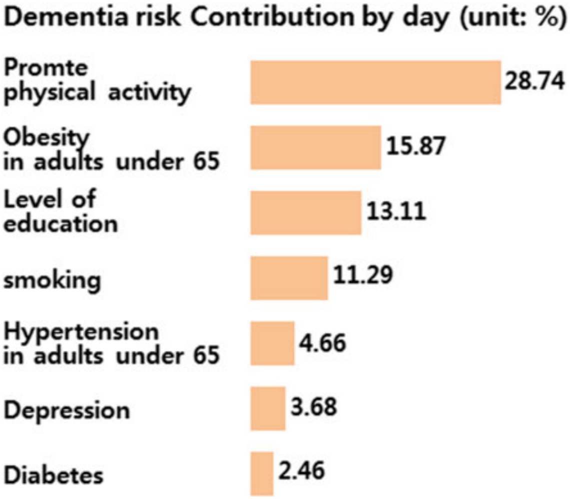
Fig. 1 Flowchart showing pathway for diagnosis and subtypes of mild cognitive impairment [39]

memory, or experiences a decrease in other cognitive functions despite absence of memory problem, or has changed in personality, or has become slow in thinking or actions, or gives an impression to others that he/she has changed [8]. MCI is classified as amnestic MIC and non-amnestic according to memory impairment [9]. Amnestic MIC, the most frequent, refers to a state in which a person can maintain his/her daily activities normally despite memory disorder, and occurs most commonly with 10–15% of cases progressing to Alzheimer's disease. Non-amnestic MCI is characterized by dysfunctions in areas other than memory, such as directional/spatio-temporal functions, executive functions, or language functions [8] (Fig. 1).

Among the 7 risk factors of dementia, such as body inactivity, obesity, education level, smoking, hypertension, depression and diabetes, 'inactivity' was the most likely to contribute to dementia (Fig. 2).

In other words, elimination of 'body inactivity' risk would have a significant effect on reducing the risk of dementia [10, 11]. Physical activity plays a role in improving cognitive function and also serves an important role in mitigating and preventing the decrease in cognitive dysfunction arising from mild cognitive impairment and aging, and it has been found that improvement in cognitive function is observed more frequently in older elderly than in relatively young elderly [12].

Fig. 2 Risk factors of dementia [11]



3 Mobile/Wearable Device and Physical Activity Platforms

IT technology has been applied to various rehabilitation areas, and various mobile/wearable devices and technologies applicable to the realm of mild/cognitive impairment which precedes the stage of dementia are emerging. A product called ‘Footlogger’ has been developed, which can not only predict the fall accidents by detecting the user’s weight distribution in the whole shoe inserts and sole but also predict the dementia by analyzing the change of gait. Also, products that take the shape of shoes or wrist watches and can keep track of the location of the elderly have appeared. Wearable type products equipped with GPS and mobile communication modems cause less concern about loss than cellular phones and are easy to carry, increasing the rate of utilization. In Japan, a service is provided by using the ‘Pepper’, a humanoid robot, as a platform to prevent dementia. Pepper can be programmed to perform various functions, such as the functions to converse naturally with customers on topics related to daily lives and family, to video-chat with family members far away, and furthermore, can communicate with nurses and physicians in real time via communication functions, thus easing the daily lives of elderly people with mild cognitive impairment [13].

Vigorous efforts need to be made for management of the elderly with mild cognitive impairment based on formation of efficient platform structure which helps increase physical activity, which is the main factor preventing the progression to dementia (Fig. 3).

The data on the amount of physical activity and basic health data (heart rate, gait patterns, etc.) of the elderly with mild cognitive impairment are stored at any time even when they are not entered personally based on sensors installed to measure the amount of exercise and utilization of wearable devices capable of measuring the

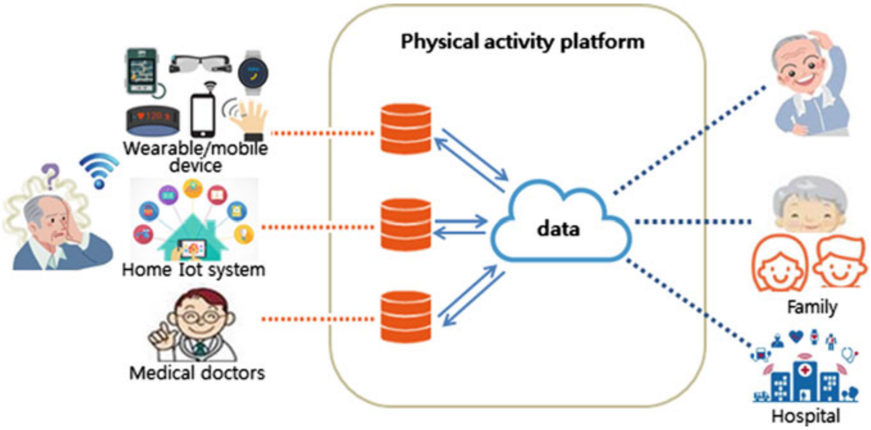


Fig. 3 Physical activity platform for the elderly with mild cognitive impairment

amount of the exercise, and furthermore, notification service can be provided to alert the users to any possible fall accident or any problem related to health conditions.

Physical activity information and health data are provided to the family through smartphones, allowing the elderly and event their family members to keep track of current health state of the elderly with mild cognitive impairment. When the amount of physical activity falls below the standard range, feedback is provided to induce the elderly to participate in physical activity by providing them with alerts and missions for prompting physical activity. When the amount of physical activity is not being monitored, alerts can be transmitted instantly in order to cope with possible emergency. In addition, one mission can be provided each day, which can be easily carried out by the elderly with mild cognitive impairment in their daily lives, leading them to experience many different physical activities.

4 Conclusion

This study proposed a physical activity platform which could induce an increase of physical activity, one of the important factors slowing down the rate of decline in cognitive functions among the elderly with mild cognitive impairment. The data on physical activity and health, collected from the sensors detecting the amount of exercise, mobile devices and wearable device outfitted with exercise sensors, are periodically updated through the platform to provide information to the elderly and their families in an open and bi-directional way. Moreover, the two-way information which is provided based on feedback process can help increase the physical activity, enabling prevention of decline in cognitive functions and promoting maintenance of cognitive functions. It would be necessary to develop the platform services capable of customized physical activity provision process based on the analyses of big data on physical activities effective for those with dementia and mild cognitive impairment in the period ahead.

References

1. Korea Statistics (2013) Cause of death statistics 2012. Daejeon, Statisrics Korea
2. Zygouris S et al (2015) Can a virtual reality cognitive training application fulfill a dual role? Using the virtual supermarket cognitive training application as a screening tool for mild cognitive impairment. J Alzheimer's Dis 44(4):1333–1347
3. Morrison GE, Simone CM, Ng NF, Hardy JL (2015) Reliability and validity of the neurocognitive performance test, a web-based neuropsychological assessment. Front Psychol 6:1652
4. Jung EY, Eun SJ, Park DK (2017) Development of evaluation program for cognitive for elderly personalized services. J Digit Art Eng Multimed 4(1):85–93
5. Korea Internet & Security Agency (2017) 2016 Survey on the internet usage
6. Jung SH (2015) Use and task of health care service for health care of elderly person and chronic ill person. Korea Insurance Research Institute
7. Jang IS Mild cognitive impairment, the next station is dementia? Preventing symptoms from worsening with early screening and treatment. The Kyunghyang newspaper. http://news.khan.co.kr/kh_news/khan_art_view.html?artid=201511041604172
8. Hong YJ 99 Dementia stories. Korean Dementia Association
9. Oh ES, Lee AY (2016) Mild cognitive impairment. J Korean Neurol Assoc 34(3):167–175
10. Petersen RC (2004) Mild cognitive impairment as a diagnostic entity. J Intern Med 256:183–194
11. Jung YH, Ko SJ (2017) Analysis of contributory factor of dementia risk factor and search for dementia management plan. Korea Institute for Health and Social Affair
12. Kim YS (2014) Physical activity and mental health. Hanyang Med Rev 34:60–65
13. KOTRA overseas market news (2017) Super-aged society Japan's dementia prevention business latest trends

Part III
Multimedia and Visualization

A Simplified Emulation Scheme for OpenGL SC 2.0 Rendering over OpenGL Embedded Systems 2.0

Nakhoon Baek

Abstract OpenGL (Open Graphics Library) is one of the most widely-used API (application programming interface)-level 3D graphics libraries. Recently, its new safety-critical profile, OpenGL SC (Safety Critical profile) 2.0 is released. To provide these new features, we design a simplified rendering scheme for emulating OpenGL SC 2.0 over OpenGL ES (Embedded System) 2.0. Since OpenGL ES 2.0 is widely used with desktops and mobile devices, our emulation can be used with wide range of graphics devices. Our new emulation scheme shows an efficient architectural way of providing all the rendering features. Prototype implementations are also presented.

Keywords Emulation · OpenGL · Safety critical profile

1 Introduction

The OpenGL library is one of the most widely used 3D graphics libraries. Actually, OpenGL means a set of standard specifications, each of which are carefully specified for its own design goals, with the common purpose of 3D rendering graphics output on the various devices [1–8].

The original OpenGL libraries are designed to support desktop PC's and workstations. Mainframes and super computers are also supported. For mobile devices and embedded systems, they introduced the new OpenGL ES series. Due to the great

N. Baek (✉)
School of Computer Science and Engineering, Kyungpook National University,
Daegu 41566, Republic of Korea
e-mail: oceancru@gmail.com

N. Baek
Software Technology Research Center, Kyungpook National University,
Daegu 41566, Republic of Korea

N. Baek
dassomey.com Inc, Daegu 41566, Republic of Korea

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_14

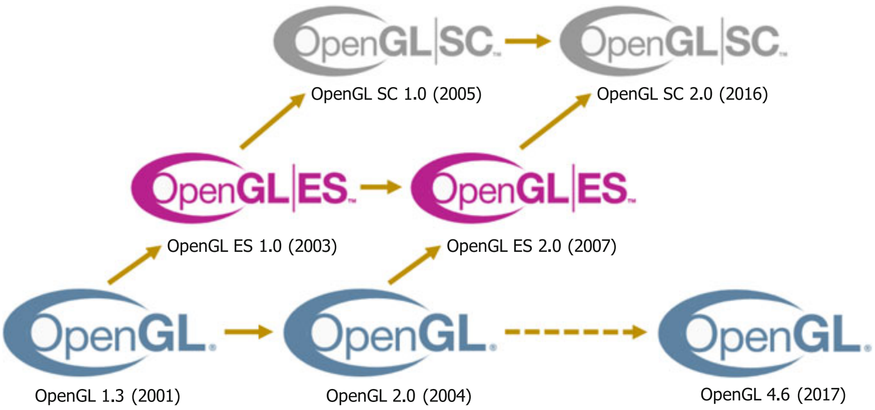


Fig. 1 The history of OpenGL SC 2.0 development

success of mobile phones including iPhones and Android phones, the OpenGL ES is now one of the most widely used graphics libraries [3, 4, 6].

OpenGL SC (OpenGL for Safety Critical) is conceptually a safety critical variation of the famous OpenGL standard. This graphics API library is designed to meet the needs of safety critical markets for avionics, industrial, military, medical and automotive applications. In the case of safety-critical markets, OpenGL SC plays the major role for the graphical interfaces. The need for this 3D graphics API is rapidly increasing with the growth of the safety-critical market [9, 10]. For the medical and automotive applications, consumer electronics markets start to strongly need this standard [11].

In the year of 2015, the Khronos Group, the fundamental standard management body of the OpenGL family, established the new OpenGL SC 2.0 specification [8]. It originally aims to a safety critical subset of OpenGL ES 2.0 [3], as shown in Fig. 1. The Khronos Group is also developing cross-API guidelines to aid in the development of open technology standards for safety critical systems.

Our goal is to provide OpenGL SC 2.0 features, as an emulation layer over OpenGL ES 2.0. Since OpenGL ES is so widely used, this emulation enables us to provide OpenGL SC 2.0 features to many feasible graphics devices. Design details and our prototype implementations will be shown in the following sections.

2 Emulation Scheme

To emulate the OpenGL SC 2.0 over OpenGL ES .20 library, we should start with analyzing and comparing the details of the API functions in both of OpenGL SC 2.0 and OpenGL ES 2.0. Those comparisons are partially performed in our previous publications of [12].

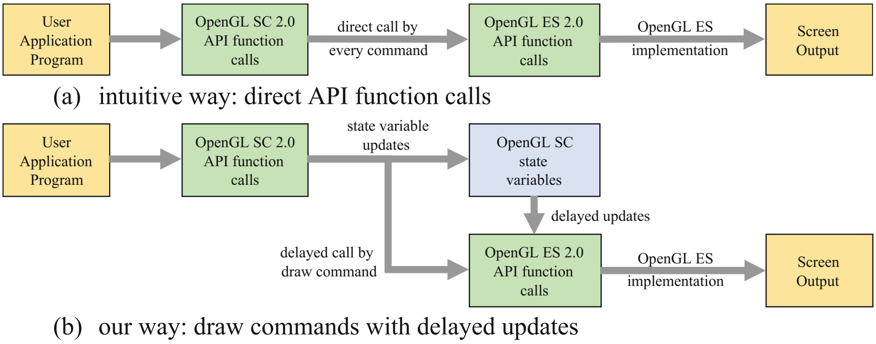

User Application Program
OpenGL SC 2.0 API function calls
direct call by every command
OpenGL ES 2.0 API function calls
OpenGL ES implementation
Screen Output
(a) intuitive way: direct API function calls
User Application Program
OpenGL SC 2.0 API function calls
state variable updates
OpenGL SC state variables
delayed updates
delayed call by draw command
OpenGL ES 2.0 API function calls
OpenGL ES implementation
Screen Output
(b) our way: draw commands with delayed updates

Fig. 2 Two ways of designing the OpenGL SC 2.0 emulation library over OpenGL ES 2.0 implementation

There have been several cases of implementing graphics emulation libraries over the other 3D graphics API libraries [11, 13–15]. In the most of these emulation library implementations, the emulation of each API functions are performed by executing or emulating the corresponding API function directly, just for each API commands, as shown in Fig. 2a.

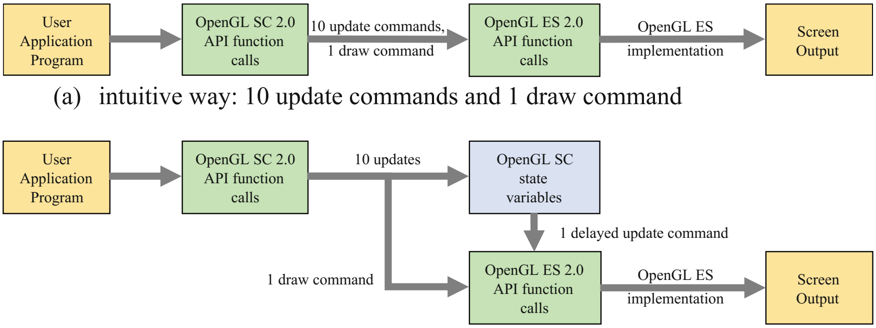
In contrast, our implementation fully uses internal state variables. Basically, OpenGL can be regarded as a state machine, whose state variables are modified by OpenGL API functions. Some special drawing commands actually execute the rendering operations, with the previously specified state variables. Our implementation fully adopt the OpenGL standard specifications. Most of the OpenGL SC API functions will perform the state variable update commands. When the drawing function is executed, we send the real drawing command to the underlying OpenGL ES implementation, with the current state variables, as shown in Fig. 2b.

Our new emulation scheme with internal state variables can show better performance, and also design efficiency with various cases of underlying systems and devices. As an example, we use the test case of Fig. 3. In a typical OpenGL programs, the user can update a single variable repeatedly, prior to the actual rendering command, as shown in Fig. 3.

In the previous emulation scheme, the emulation library will execute the update command and the drawing command exactly after the API function calls, as shown in Fig. 4a. In our case, the update commands are delayed and executed only one, just before the drawing command, as shown in Fig. 4b. Hence, for the example program of Fig. 3, our new scheme can reduce the total executed commands to 2, in comparison with the total of 11 in the previous schemes.

```
// 10 update commands
glColor4f( 0.1f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.2f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.3f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.4f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.5f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.6f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.7f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.8f, 0.0f, 0.0f, 1.0f );
glColor4f( 0.9f, 0.0f, 0.0f, 1.0f );
glColor4f( 1.0f, 0.0f, 0.0f, 1.0f );
// 1 draw command
glDrawArrays( GL_TRIANGLES, 0, 3 );
```

Fig. 3 An example scenario: the application program code segment



(a) intuitive way: 10 update commands and 1 draw command

(b) our way: 1 delayed update command and 1 draw command

Fig. 4 An example scenario: 10 duplicated update commands and 1 draw command case

3 Conclusion

With a newly released graphics library, we need a fast prototyping of all the features, and one of the simplest way is providing an emulation library. In this paper, we aimed to provide the OpenGL SC 2.0 emulation library over the OpenGL ES 2.0. Our focus is the delayed state update features. This new emulation scheme was fully tested and implemented with our current prototype implementation of OpenGL SC 2.0 emulation library.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant 2016R1D1A3B03935488).
This study was also supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005).

References

1. Segal M, Akeley K (2006) The OpenGL graphics system: a specification, version 2.1
2. Blythe D (2008) OpenGL ES common/common-lite profile specification, version 1.1.12 (Full Specification)
3. Munshi A, Leech J (2010) OpenGL ES common profile specification, version 2.0.25 (Full Specification)
4. Lipchak B (2013) OpenGL ES. Version 3.0.2
5. Kessenich J (2006) The OpenGL shading language, language version: 1.20
6. Simpson RJ (2013) The OpenGL ES shading language, language version: 1.00
7. Stockwell B (2009) OpenGL SC: safety-critical profile specification, version 1.0.1. Khronos Group
8. Fabius A, Viggers S (2016) OpenGL SC, version 2.0.0 full specification.
9. Cole P (2005) OpenGL ES SC—open standard embedded graphics API for safety critical applications. In: 24th digital avionics systems conference
10. Mark S (2005) Solving the embedded OpenGL puzzle—making standards, tools, and APIs work together in highly embedded and safety critical environments. In: 24th digital avionics systems conference
11. Baek N, Baeck GJ (2010) Design of OpenGL SC emulation library over the desktop OpenGL 1.3. In: 29th digital avionics systems conference
12. Baek N An emulation scheme for OpenGL SC 2.0 over OpenGL. J Supercomput (submitted)
13. Baek N, Lee H (2012) A cost-effective OpenGL SC solution for the consumer electronics market: an emulation library approach. In: IEEE international conference on consumer electronics digest of technical papers
14. Baek N, Lee H (2011) Implementing OpenGL SC over OpenGL 1.1+. In: IEEE international conference on consumer electronics digest of technical papers
15. Baek N, Yoo K-H (2015) Emulating OpenGL ES 2.0 over the desktop OpenGL. Clust Comput 18(1):165–175

Robust Sound Localization Algorithm for Intelligent Acoustic Surveillance System

Kyusik Park

Abstract A new sound localization algorithm for intelligent surveillance system is proposed in this paper. This study mainly focuses on the building of the reliable tetrahedron-shaped microphone array that allows noise robust sound localization. The proposed system is derived from the traditional triangular microphone array and developed using a plane coordinate transformation method. A real-time sound localization experiments with a proposed method, the previously known triangular microphone array and existing tetrahedral microphone array system are conducted and compared. The experimental results show that the proposed system outperforms all the comparing systems in terms of the detection rate of the sound source. The proposed system also confirms the robustness of the noise effect in low SNR environment.

Keywords Sound localization · Intelligent acoustic surveillance
Coordinate transformation

1 Introduction

Sound source localization that tracks the direction of a sound source in arbitrary space is a fundamental and practical research topic with many applications. The technology can be applied to various systems such as teleconference system, audio/visual communication system, robot recognition system, artificial intelligence system, etc. It is also applicable to intelligent surveillance service that detects acoustic event and point camera to the event, in which the demand for the technology is increasing rapidly [1].

There are three main sound source localization methods in the literature: beamforming [2, 3], MUSIC (MUltiple SIgnal Classification) [4, 5], and TDOA (Time

K. Park (✉)
Department of Software, Dankook University, Suji-gu, Yongin-si,
Gyenggi-do, South Korea
e-mail: kspark@dankook.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_15

Delay Of Arrival). Beamforming steers the sensor array to search for peak amplitude in output signal. MUSIC method estimates the frequency content of a signal or autocorrelation matrix using an eigenspace method from the multiple microphones. TDOA is a most popular method for the source localization. It measures the time delay of arrival (TDOA) among the microphone array and estimates the source location using the geometric relation of a sound source and the microphones. In general, two microphones are needed to estimate the direction of sound source in plane. On the other hand, at least three microphones are required to localize a sound source in a plane, and four microphones required for locating sound source in 3D space.

The most important part of TDOA method is how accurately measures the time delay of arrival of sound source from the microphone sensor array. A classical approach to calculate TDOA is a Cross-Correlation (CC) method. CC method analyzes the cross-correlation of the signals received by a pair of microphone and search for the peak amplitude as a time-delay [6]. It is easy to understand, but it is quite sensitive to environmental noise. For this reason, the generalized cross-correlation (GCC) method was developed as an improvement of CC method [7, 8]. GCC method alleviates noise effect by giving weighting function to CC method. Depending on the weighting function, GCC method is further divided into ROTH [9], SCOT [10], CPSP [11], and Eckart Processor [12] methods. Corresponding literature reviews are followings. Feng et al. [13] proposed a TDOA-based triangulation method with four microphones of Phase Difference of Arrival (PDOA) estimation. Fan et al. [14] studied a localization of sound source and distance by plane microphone array. They introduce a quasi L1-autocorrelation algorithm and an interpolation algorithm for improving estimation accuracy. Hamada and Ozeki [15] proposes the method for sound localization using tetrahedral microphone array. They achieved efficient 3D DOA (Direction Of Arrival) estimation using spherical pseudo-histogram and weighted past histograms. Lee and Choi [16] suggests a spherical sound source localization method using triangular microphone array. Peng et al. [17] proposed an acoustical localization system using microphone array for mobile robot. They used a regular tetrahedral microphone array to estimate azimuth angle of the sound source. On the other hand, in Ref. [18], they provided a triangular microphone array to calculate the azimuth and elevation angle of a sound source by assuming plane sound wave.

This paper is a further extension and elaboration of the works in [18]. In their works, they setup three-microphone array of equilateral triangular form to estimate the directional vector of the sound source, so called unit sighting vector. From the geometrical relationship between the microphone array and the sound source, the azimuth and elevation angle of the sound source was derived with a plane wave assumption. In contrast to work in [18], this paper mainly focuses on the building of tetrahedron-shaped microphone array that allows noise robust sound localization system. This can be achieved by adding one microphone to the triangular microphone array in [18] and the use of a plane coordinate transformation. A main target for the proposed system is the sound localization system applicable to intelligent CCTV surveillance system that covers 5–10 m distance range under the noise environment.

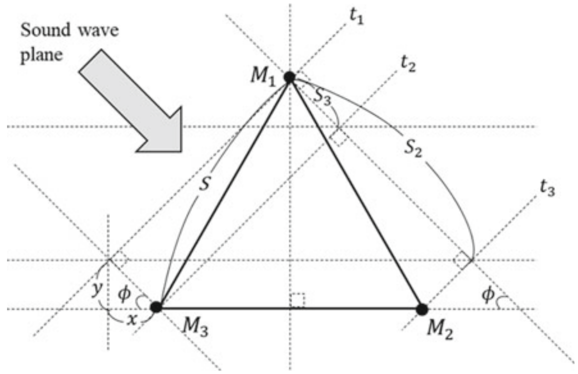


Fig. 1 Sound localization system using three-microphone array of equilateral triangular form

The rest of the paper is organized as follows. Section 2 reviews previous works done in [18]. This section also describes the proposed method. Section 3 evaluates and demonstrates potential usefulness of the proposed method through real-time experiment. Finally, a concluding remark is given in Sect. 4.

2 Sound Localization Algorithm

2.1 Sound Localization with Triangular Microphone Array

This section summarizes the sound localization algorithm described in [18]. It will be used as a basis algorithm to build proposed system developed in Sect. 2.2. Figure 1 shows sound localization system to estimate the directivity of the sound source using three-microphone array. It is a graphical representation of how the sound wave plane passes through the microphone array in time sequence t_1, t_2, t_3. Here, the sound source is assumed to reach a microphone as a plane wave. This is a reasonable assumption if the sound wave is generated from some reasonable distance from a microphone sensor [18].

In this figure, three microphones form an equilateral triangle. Here, M_1, M_2, M_3 denote vector coordinate of three microphones and a coordinate origin is placed at M_3. The sound source wave is assumed to hit M_1 first, M_3 second, and M_2 last. In this figure, S is a distance between the microphones. S_2 is a distance where a plane wave travels from M_1 to M_2 and S_3 is a distance where a plane wave travels from M_1 to M_3. Note that the distance S_2, S_3 can be represented in terms of TDOA (Time Delay Of Arrival) between the microphone pair such as $S_2 = TDOA_{12} \cdot v$ and $S_3 = TDOA_{13} \cdot v$ where $v$is sound speed.

Furthermore, from the figure, the distance S_2 and S_3 can be derived as follows

$$S_2 = S \cdot \cos(60 - \theta) = S \cdot \left(\frac{\sqrt{3}}{2} sin\theta + \frac{1}{2} cos\theta\right)$$
$$S_3 = S \cdot \cos(120 - \theta) = S \cdot \left(\frac{\sqrt{3}}{2} sin\theta - \frac{1}{2} cos\theta\right) \quad (1)$$

using triangular geometry $\cos(\alpha \pm \beta) = \cos\alpha \cdot \cos\beta \mp \sin\alpha \cdot \sin\beta$.

By adding and subtract each term in Eq. (1), we will have

$$S_2 + S_3 = S \cdot \left(\frac{\sqrt{3}}{2} sin\theta + \frac{1}{2} cos\theta + \frac{\sqrt{3}}{2} sin\theta - \frac{1}{2} cos\theta\right) = S\sqrt{3} sin\theta$$
$$S_2 - S_3 = S \cdot \left(\frac{\sqrt{3}}{2} sin\theta + \frac{1}{2} cos\theta - \frac{\sqrt{3}}{2} sin\theta + \frac{1}{2} cos\theta\right) = S cos\theta \quad (2)$$

Then, the directional vector of sound source (x, y, z) can be obtained as

$$x = \frac{S_3(S_2 - S_3)}{S} \quad y = \frac{S_3(S_2 + S_3)}{S\sqrt{3}} \quad \mathrm{z} = \sqrt{S_3^2 - x^2 - y^2} \quad (3)$$

where $\cos\theta = \frac{x}{S_3}$, $sin\theta = \frac{y}{S_3}$.

The azimuth angle θ and the elevation angle φ of the plane wave of the sound source are then derived by

$$\theta = tan^{-1}\left(\frac{y}{x}\right), \quad \varphi = tan^{-1}\left(\frac{z}{\sqrt{x^2 + y^2}}\right) \quad (4)$$

2.2 Proposed Algorithm

Sound localization system estimates directional information of sound source based on TDOA (Time Delay Of Arrival) between the microphone pairs. To obtain directional information such as azimuth and elevation angle of the sound source in three dimensions, the system requires at least three microphones as shown in Fig. 1. If one of the microphones fails to detect proper signal level, the system cannot produce a reliable estimate. For this reason, this paper suggests a tetrahedron- shaped sound localization system as shown in Fig. 2. By adding one microphone to triangular-shaped system in Fig. 1 and the use of a simple coordinate transform method allows more robust and reliable sound localization system.

From the Fig. 2, let a distance between coordinate origin and the microphone mic0 is a, then the vector coordinate of the microphones can be represented as follows.

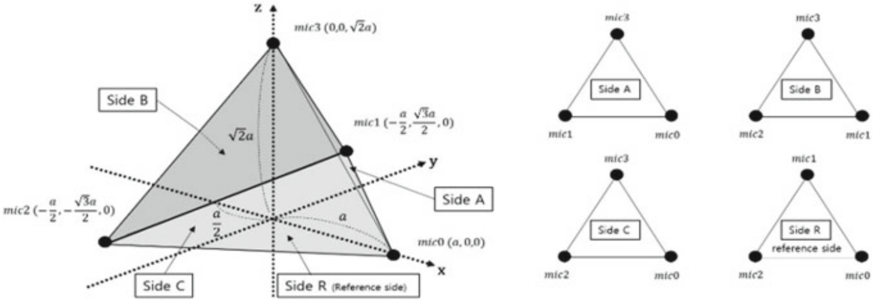


Fig. 2 Tetrahedron-shaped microphone array system with four microphones

$$mic0 = (a, 0, 0), mic1 = \left(-\frac{a}{2}, \frac{\sqrt{3}a}{2}, 0\right), mic2 = \left(-\frac{a}{2}, -\frac{\sqrt{3}a}{2}, 0\right),$$
$$mic3 = (0, 0, \sqrt{2}a) \tag{5}$$

From the Fig. 2, let a distance between coordinate origin and the microphone mic0 is a, then the vector coordinate of the microphones can be represented as follows.

$$mic0 = (a, 0, 0), \quad mic1 = \left(-\frac{a}{2}, \frac{\sqrt{3}a}{2}, 0\right), \quad mic2 = \left(-\frac{a}{2}, -\frac{\sqrt{3}a}{2}, 0\right),$$
$$mic3 = (0, 0, \sqrt{2}a) \tag{6}$$

With an extra microphone mic3, we can have four triangle faces—a reference side R (mic0-mic1-mic2), side A (mic0-mic1-mic3), side B (mic1-mic2-mic3), and a side C (mic0-mic2-mic3). Each triangle side is formed with triangular microphone array and so, the directional information (azimuth angle θ, elevation angle ϕ) of the sound source based can be estimated at each side using the algorithm introduced in Sect. 2.1. Therefore, the proposed system in Fig. 2 can have four candidate estimates of directional information from the four triangle sides. Each estimate of the directional information obtained from side A, B, C is now coordinate transformed to reference side R so that they can be compared and averaged in the interval where the directional estimates intersect each other. This method has the advantage of being robust against ambient noise or error since the final result can be derived by using the results of the remaining triangular microphone arrays even if errors occur in any of the microphones. In summary, the proposed algorithm can be implemented as follows.

1. Estimate directional information of the sound source from the four triangle faces to have four candidate estimates of the azimuth angle and the elevation angle.
2. Do coordinate transform of the estimates from the side A, B, C to the reference side R.

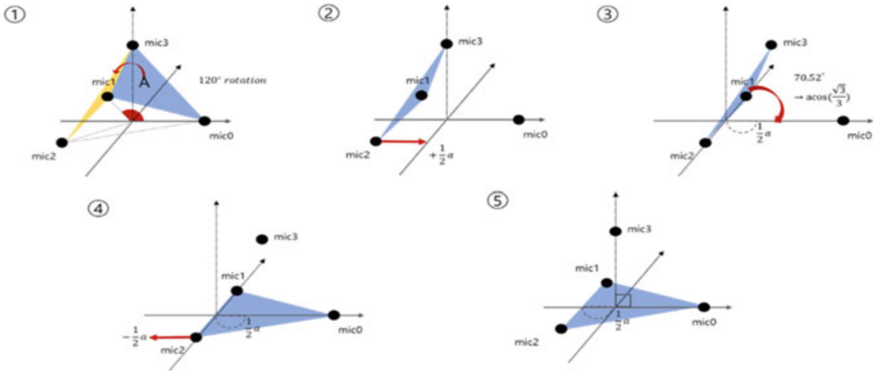


Fig. 3 Coordinate transform from a side A to a reference side R

3. Average out four candidate estimates in the interval where the directional estimates intersect each other.

Figure 3 describes five steps in coordinate transform from a side A to a reference side R.

Referring Fig. 3, Eq. (7) describes matrix equation that can transform coordinate value (x_A, y_A, z_A) in side A to the coordinate value (x_{AR}, y_{AR}, z_{AR}) in a reference side R.

$$\begin{pmatrix} x_{AR} \\ y_{AR} \\ z_{AR} \end{pmatrix} = \begin{bmatrix} \cos\left(-\frac{2}{3}\pi\right) & -\sin\left(-\frac{2}{3}\pi\right) & 0 \\ \sin\left(-\frac{2}{3}\pi\right) & \cos\left(-\frac{2}{3}\pi\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
$$\cdot \left[\begin{bmatrix} \cos\left(\frac{70.52}{180}\pi\right) & 0 & -\sin\left(\frac{70.52}{180}\pi\right) \\ 0 & 1 & 0 \\ \sin\left(\frac{70.52}{180}\pi\right) & 0 & \cos\left(\frac{70.52}{180}\pi\right) \end{bmatrix} \cdot \left\{ \begin{pmatrix} x_A \\ y_A \\ z_A \end{pmatrix} + \begin{pmatrix} \frac{1}{2}a \\ 0 \\ 0 \end{pmatrix} \right\} + \begin{pmatrix} -\frac{1}{2}a \\ 0 \\ 0 \end{pmatrix} \right] \tag{7}$$

In the same manner, the coordinate values (x_B, y_B, z_B) from side B and (x_C, y_C, z_C) from side C can be transformed to a reference side R as (x_{BR}, y_{BR}, z_{BR}) and (x_{CR}, y_{CR}, z_{CR}) respectively.

$$\begin{pmatrix} x_{BR} \\ y_{BR} \\ z_{BR} \end{pmatrix} = \begin{bmatrix} \cos\left(\frac{70.52}{180}\pi\right) & 0 & -\sin\left(\frac{70.52}{180}\pi\right) \\ 0 & 1 & 0 \\ \sin\left(\frac{70.52}{180}\pi\right) & 0 & \cos\left(\frac{70.52}{180}\pi\right) \end{bmatrix} \cdot \left\{ \begin{pmatrix} x_B \\ y_B \\ z_B \end{pmatrix} + \begin{pmatrix} \frac{1}{2}a \\ 0 \\ 0 \end{pmatrix} \right\} + \begin{pmatrix} -\frac{1}{2}a \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} x_{CR} \\ y_{CR} \\ z_{CR} \end{pmatrix} = \begin{bmatrix} \cos\left(\frac{2}{3}\pi\right) & -\sin\left(\frac{2}{3}\pi\right) & 0 \\ \sin\left(\frac{2}{3}\pi\right) & \cos\left(\frac{2}{3}\pi\right) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \left[\begin{bmatrix} \cos\left(\frac{70.52}{180}\pi\right) & 0 & -\sin\left(\frac{70.52}{180}\pi\right) \\ 0 & 1 & 0 \\ \sin\left(\frac{70.52}{180}\pi\right) & 0 & \cos\left(\frac{70.52}{180}\pi\right) \end{bmatrix} \right.$$
$$\left. \cdot \left\{ \begin{pmatrix} x_C \\ y_C \\ z_C \end{pmatrix} + \begin{pmatrix} \frac{1}{2}a \\ 0 \\ 0 \end{pmatrix} \right\} \begin{pmatrix} -\frac{1}{2}a \\ 0 \\ 0 \end{pmatrix} \right] \tag{8}$$

On the other hand, TDOA (Time Delay Of Arrival) information between the two microphone signals x_1, x_2 is measured as CPSP (Cross Power Spectrum Phase) method [11]. In other words, TDOA is chosen to be τ that maximize the Generalized Cross-Correlation (GCC) in Eq. (9),

$$R^{(g)}_{x_1x_2}(\tau) = \int\limits_{-\infty}^{\infty} \Psi_g(f) S_{z_1x_2}(f) e^{j2\pi f\tau} df \tag{9}$$

where CPSP defines general function $\Psi_g(f)$ as a reciprocal of cross-power spectrum of microphone signals such that $\Psi_g(f) = \frac{1}{|S_{x_1x_2}(f)|}$.

3 Experimental Results

Figure 4 describes overall system algorithm in block diagram. First of all, microphone signals are acquired at 25.6 kHz sampling rate using National Instrument DAQ NI-9234. These signals are processed on frame by frame basis and each frame consists of 30,000 samples. TDOA is then computed using CPSP algorithm in every frame between all possible microphone pairs. Based on TDOA estimates, azimuth and elevation angles of the sound source in reference side R, side A, side B, and side C, are calculated using the triangular microphone array described in Sect. 2.1. Then each estimate of the directional information obtained from side A, B, C is coordinate transformed to reference side R so that they can be compared and averaged to yield final result.

In order to verify the system performance, a real-time sound localization system with (1) triangular microphone array in Fig. 1, (2) the proposed tetrahedron-shaped microphone array in Fig. 2 and (3) tetrahedral microphone array in Ref. [17] are implemented and compared. In Ref. [17], a regular tetrahedral microphone array system is described, and the azimuth angle of the sound source was derived using

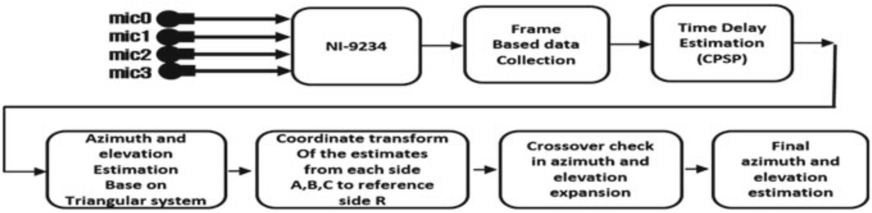

mic0
mic1
mic2
mic3
NI-9234
Frame Based data Collection
Time Delay Estimation (CPSP)
Azimuth and elevation Estimation Base on Triangular system
Coordinate transform Of the estimates from each side A,B,C to reference side R
Crossover check in azimuth and elevation expansion
Final azimuth and elevation estimation

Fig. 4 Overall proposed sound localization system

geometrical relationship of the microphones. With an assumption that the distance between the sound source and the microphone array is far greater than the size of microphone array, the azimuth angle is approximated as

$$\theta \approx tan^{-1}(\sqrt{3}\frac{TDOA_{13} - TDOA_{23}}{TDOA_{13} + TDOA_{23} - 2 \cdot TDOA_{03}}) \qquad (10)$$

where $TDOA_{ij}$ is a time-delay of arrival between microphone i and j

Experiments run on Intel Core™2 Duo T8300 2.4G CPU; triangular localization system consumes 0.32 s of computation time per each frame while both tetrahedron localization system required an average of 0.62 s. Most of the computational operations were taken to the time-delay estimation algorithm. A total of 240 s long acoustic data was prepared for the experiment; they are randomly combined to evaluate the sound localization performance. A sound source was placed at a distance of 5 m, 10 m respectively. Environmental SNR was measured at each distance without sound source present and they were 8 dB at 5 m and 2 dB at 10 m, respectively. At each distance placement, sound source was set to 20°, 40°, 60°, and 80° in elevation angles, and the azimuth angle at each elevation angle was set between 30° and 330° in 30° intervals. For the system performance evaluation, a detection ratio is computed. If the estimated azimuth and elevation angles from the experiment are within ±10° compared to ground truth, we treat this case as a successful detection; otherwise it was regarded as a detection failure for the error range beyond the ±10°.

Figure 5 compares the experimental results for (1) triangular microphone array in Fig. 1 (Triangle), (2) the proposed tetrahedron-shaped microphone array in Fig. 2 (P_Tetrahedron) and (3) tetrahedral microphone array in Ref. [17] (Q_Tetrahedron). The sound source is placed at distance 5 m and 10 m respectively with elevation angle 20°, 40°, 60°, and 80°, and the azimuth angle at each elevation angle between 30° and 330° in 30° intervals. The experiment is repeated for every 200 frames of sound source to derive the detection rate. From the figure, as the distance decrease from 10 to 5 m, the better overall system performance to all the systems. Especially, we note that the higher localization performance of the proposed method is attained for all distance and directions. In terms of the measuring distance from sound source to the microphone array, the proposed system (P_Tetrahedron) is superior to triangle microphone system (Triangle) by 5% improvement at distance 5 m, and 9% improvement at distance 10 m in average. When comparing the proposed system (P_Tetrahedron)

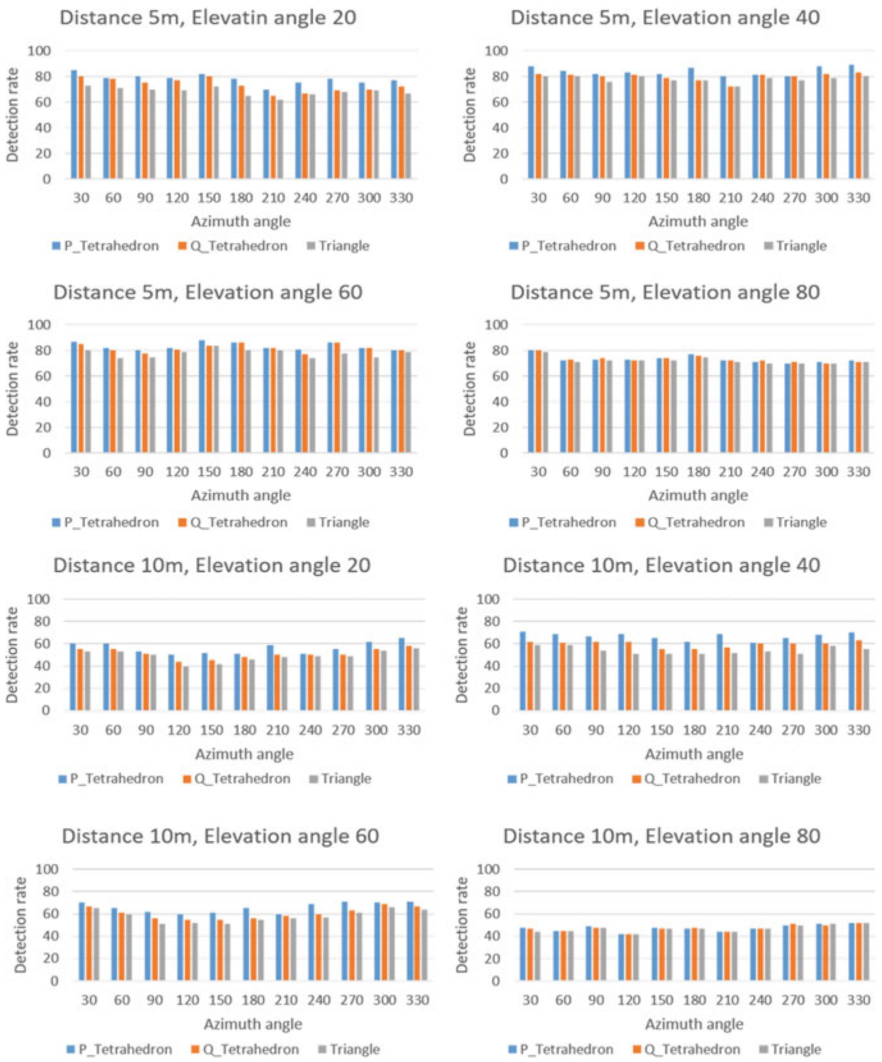


Fig. 5 Performance comparison at distance 5 and 10 m

and tetrahedral microphone array in Ref. [17] (Q_Tetrahedron), the proposed system shows about 5% better performance in average in all distance cases. As seen from the figure, the detection rate become smaller when the measuring distance is farther from 5 to 10 m. This is because when the sound source gets farther, the TDOA's calculation error between the microphones will be more evident so that the resulting detection rate will be less accurate. The detection rate is satisfactory when the measuring distance is 5 m. However, at the distance 10 m, the system performance

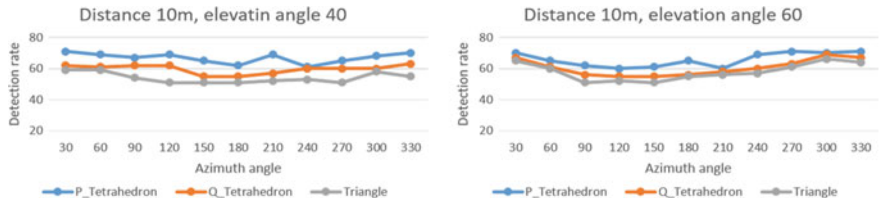


Fig. 6 Performance comparison at distance 10 m

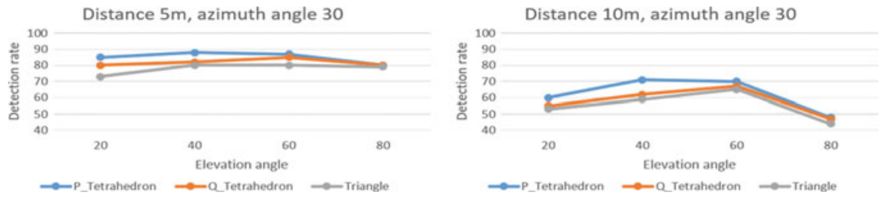


Fig. 7 Performance comparisons as the distance varying

is highly dependent on the elevation angle of the sound source. At a distance 10 m and elevation angle 80°, the detection rate is unacceptable as below 60%.

Figure 6 compares system performance of the comparing systems at distance 10 m. We note that the proposed system demonstrates the robustness of the noise effect in low SNR environment. The proposed system is quite comparative to "Triangle" and "Q_tetrahedron". The proposed system outperforms the comparing systems about 4–7% at distance 5 m, and 7–16% at distance 10 m. Furthermore, at distance 10 m, azimuth angle 120°, and elevation angle 40°, the proposed algorithm outperforms the triangle microphone system about 16% in maximum.

From Fig. 7, near the edge of the elevation angle such as 0° and 80°, all comparing systems shows worse detection rate. Especially at elevation angle 80°, the detection rate is the worst and there was no significant difference in performance between the three algorithms. This is because, in all systems, the time-delay estimation error increases as the elevation angle approaches to the edges. On the other hand, the system performance of all three systems is higher and not much change as the elevation angle increase between the edges.

4 Conclusion

In this paper, we study a robust sound localization system applicable to intelligent CCTV surveillance system. We constructed a tetrahedron-shaped microphone array with one microphone added to existing triangular microphone system and a coordinate transform method. We propose a novel sound localization system that is robust to ambient noise and errors. Through a real-time comparison experiments, the proposed system outperforms all the comparing systems in terms of the detection rate of the sound source. Especially, the proposed system demonstrates the robustness of the noise effect in low SNR environment. In future studies, we will use the TDOA and the interaural level difference (ILD) of the sound source arriving at each pair of microphone as source localization cue. In addition, the efficient system implementation for more robust sound localization system will be studied.

Acknowledgements This work was supported by the Technology Development Program S2417278 funded by the Ministry of SMEs and Startups (MSS, Korea).

References

1. Park S, Cho M, Park Y (2014) Development of the intelligent CCTV based management system for tracing community safety risk 16(1):50–56. National Disaster Management Institute, National Disaster Management (In Korean)
2. Zhang C, Florêncio D, Ba DE, Zhang Z (2008) Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans Multimed 10(3)
3. Valin J-M, Michaud F, Hadjou B, Rouat J (2004) Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. Proc ICRA 1:1033–1038
4. Johnson DH, Dedgeon DE (1993) Array signal processing. PTRP Prentice Hall
5. Ishi CT, Chatot O, Ishiguro H, Hagita N (2009) Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In: The 2009 IEEE/RSJ international conference on intelligent robots and systems, St. Louis, USA, 11–15 Oct 2009
6. Papoulis A (1966) Probability random variables and stochastic process. McGraw-Hill, New York
7. Knapp CH, Carter GC (1976) The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process ASSP-24:320–327
8. Qinqin Z, Linghua Z (2015) Study of delay estimation in acoustic source localization based on microphone array. In: Proceedings of IAEAC, China
9. Roth PR (1971) Effective measurements using digital signal analysis. IEEE Spectr 8:62–70
10. Carter GC, Nuttal AH, Cable PG (1973) The smoothed coherence transform. Proc IEEE 6:1497–1498
11. Omologo M, Svaizer P (1994) Acoustic event localization using a cross-power spectrum phase based technique. In: Proceedings of ICASSP'94, pp 273–276, Adelaide, Australia
12. Eckhart C (1952) Optimal rectifier systems for detection of steady signals, Scripps. Inst Oceanography, Marine Physical Lab. Univ. California, Rep. SIO 12692, Ref, 52–11
13. Feng M, Diange Y, Rujia W, Junjie W, Ziteng W, Xiaomin L (2014) A triangulation based on phase difference of arrival estimation for sound source localization. In: The 21st international congress on sound and vibration, pp 13–17, Beijing, China
14. Fan J, Luo Q, Ma D (2010) Localization estimation of sound source by microphones array. In: Symposium on security detection and information processing. Elsevier, pp 312–317
15. Ozeki K, Hamada N (2006) Estimating directions of multiple sound sources using tetrahedral microphone array. In: IEEE region 10 conference
16. Lee B, Choi J (2009) Spherical localization of sound source using triangular microphone array. In: Proceedings of KACC2009, pp 360–363 (in Korean)
17. Peng Y, Hao S, Zu L (2009) An acoustic localization system using microphone array for mobile robot. Int J Intell Eng Syst 2(4):18–26
18. McNelis N, Conner N (1993) Methods and apparatus for determining the trajectory of a supersonic projectile, US Patent 5,241,518

A Multiple Optical Vision System of Virtual Display

Yongseok Chi

Abstract This paper proposes a head-up display that displays virtual images of horizontal multichannel windows in order to protect drivers from potential dangers and possible accidents, which occur due to eye recognition response time for object distance in fast moving vehicles. A head-up display with a combiner lens structure that can simultaneously project multiple virtual images without antireflective coating material applied to the windshield has been studied. This includes an optical ray tracing system, which displays a virtual image of more than 8 inches in size at a distance of more than 3 m from the human eye. In addition, asymmetric and aspheric optical folding mirrors and a concave mirror system are applied to enlarge the virtual image. The study evaluation was realized through the optical distortion characteristics and the brightness uniformity of two lighting systems. The brightness uniformity of the two virtual displays exhibited high performances of 83.4 and 82.2%, and the optical distortion is reduced to less than 5% of the TV distortion.

Keywords Combiner system · Optical lens · Multiple micro panels

1 Introduction

The human eye recognizes the size, shape, surface structure, gloss, transparency, and color of an object. In addition, the eye has characteristic imaging factors, photosensitivity factors, light and dark adaptation, and chromatic adaptation. The focal length of the human eye adapts proportionally to time whereas the adaptability of the eye, based on the object distance varying in time, depends on the distance of the object from the eye as well as the age and sex of the human. This adaptation time can lead to lethal accidents for pilots and drivers. Head-up or down displays have been used to display virtual images that deliver various information to the pilots of civil and military airplanes at a certain distance from the eye [1]. The need to both ensure the

Y. Chi (✉)
Division of Mechatronics Engineering, Dongseo University, Busan, South Korea
e-mail: ys.chi@dongseo.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_16

safety of the airplane and evaluate surrounding situations makes it difficult for pilots to keep their field of vision directed towards a far distance through the windshield while simultaneously keeping an eye on the information shown by the instrument panel located close to the eyes [2].

In order to resolve such difficulties, virtual images have been configured to reduce the difference in distance between distant objects and the instrument panel information to allow pilots to be easily aware of the frequently used information. In addition, head-up display (HUD) technology has been developed that can overlap real objects with relevant system information [3]. Car manufacturers have adopted this technology and installed head-up displays that can provide various information about the automobile (speed, warnings, mini maps) via virtual images for the safety and convenience of the drivers [4].

The built-in head-up display in automobiles is designed to deliver limited content such as automobile information in mono colors, and hence, cannot completely accommodate the diverse needs of different drivers. The advancement of smartphones has provided a broad range of content like navigation information and additional features; however, there are technical limits of displaying such diverse image information in automobiles. Moreover, the limited space within the automobiles and not being able to increase the image height cause difficulties to the drivers in easily identifying the image information of the head-up displays.

In order to solve several technical problems of the existing windshield-type head-up displays, a combiner lens and two liquid crystal display (LCD) panels, 0.47 inches in size, were used to enlarge the virtual image, resulting in a head-up display with a multichannel image display. The structure of the multichannel head-up display has been modified from the existing head-up display to display multiple channels by enlarging the display in the horizontal direction [5]. The monochrome limitation of the head-up display was overcome by utilizing high-luminance light-emitting diodes (LEDs) to produce 24-bit colors of red, green, and blue. LCD panels and environment-friendly high-luminance LEDs have the advantage of improving brightness and contrast such that the drivers can easily distinguish objects from the image information [6, 7].

A comparison between the windshield-type head-up display and combiner-lens-type is described in the body of Sect. 2. Section3 presents the head-up display system with a multichannel virtual image display configured as a combiner-type in which two LCDs with a size of 0.47 inches are used. The experiment evaluation and analysis are described in Sect. 4. Sections 4 and 5 describe the evaluation of the configured technology as well as the conclusions.

2 Analysis of Head-up Display

The head-up display built in existing automobiles projects images by applying antireflective coating inserted into the windshield glass. The problems associated with this are that the system is bulky, overly complex, and limited by the engine temperature

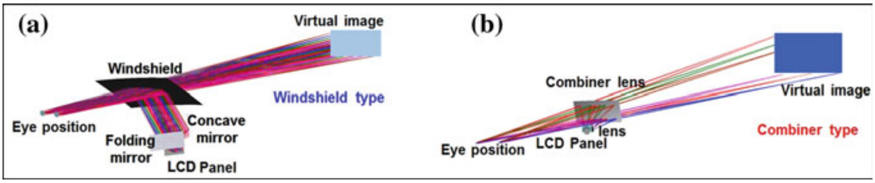


Fig. 1 The windshield-type (**a**) and combiner-type projection lens (**b**) optical ray tracing of head-up display system

and space. The most critical problem is the difficulty in delivering diverse information due to the small size of the virtual image and limited content. On the other hand, the combiner-lens-type can resolve spatial constraints; however, it has limitations in enlarging the image.

Figure 1 shows the comparison of the ray tracing between the windshield- and combiner- types designed to display virtual images 8 inches in size at a certain distance from the human eye. The windshield-type ray tracing employs antireflective coating that enables reflection and penetration to allow the human eye box to recognize the off-axis image projected through an asymmetric and aspheric folding mirror or a concave mirror from an LCD panel with a size of 18 inches. The combiner-type ray tracing of a combiner-lens-type head-up display allows virtual images to be displayed in multichannel without the windshield [8].

The design of the combiner-lens-type head-up is based on the design of an aspheric transmission-type optical lens, where lenses with different reflection angles are used at the optical surface on the front and rear of the lens to remove the ghost image [9]. The reflection ratio of the combiner lens is configured to be between 20 and 35% by applying less than 0.5% of antireflective coating for the maximum brightness of the system and obscuration removal of the image. Backward ray tracing from the virtual object to the display panel is used for optical imaging design, which minimizes the optical distortion value required for designing multi-zoom that enlarges the virtual image.

The location of the eye box is aspherically designed to cover the entire area by changing from one-dimensional displacement to two-dimensional. The field of view (FOV) and range of the eye motion box (EMB) are defined as follows.

$$\mathrm{EMB} = \mathrm{D} - (\mathrm{LE} \times \mathrm{S})/\mathrm{F} \tag{1}$$

$$\mathrm{FOV} = 2 \times \arctan(\mathrm{S}/2\mathrm{F}) = 2 \times \arctan(\mathrm{D}/2\mathrm{LE}) \tag{2}$$

where $\mathrm{D} \ll (\mathrm{S}/\mathrm{F})$.

Eye motion box (EMB is the range of movement of the eye box, D is the diameter of the lens, LE is the distance from the eye to the lens, S is the size of the virtual image display, F is the focal length of the lens, and FOV is the field of view. The ranges of the FOV and EMB are determined by the diameter of the lens and the distance between the eye and the object, respectively. If the distance from the lens to

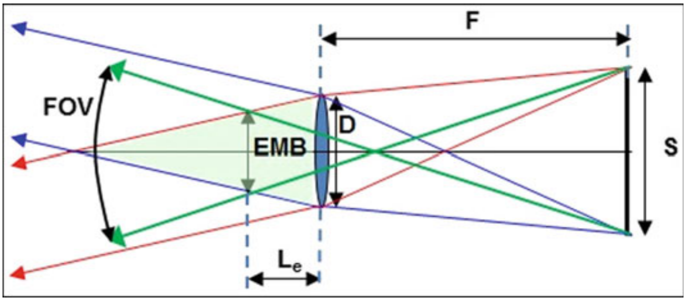


Fig. 2 Geometry of the eye motion box (EMB) and the field of view (FOV) of the HUD

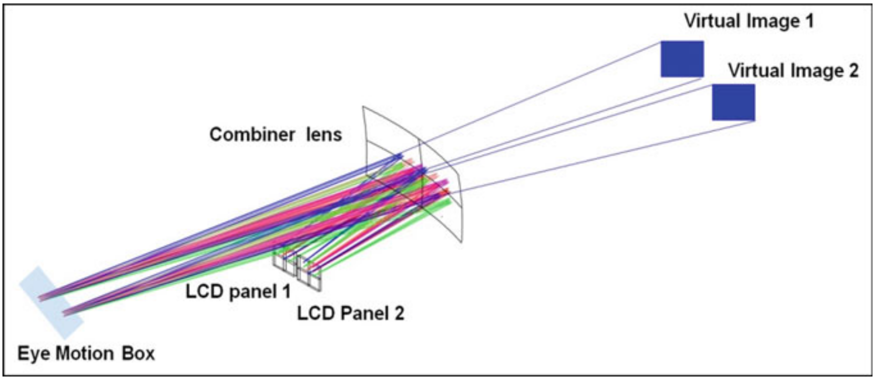


Fig. 3 The optical ray tracing design of the HUD system with a combiner lens for multi windows

the eye increases, the movement ranges of the FOV and EMB decreases, as shown in Fig. 2. In other words, the EMB and FOV are proportional to the diameter of the lens. Hence, to increase the FOV, either the diameter of combiner lens should be increased or the distance from the eye to the lens should be reduced [10].

3 Optical Combiner Type Head-up Display for Virtual Image

The ratio of horizontal and vertical display is denoted as H (horizontal axis) and V (vertical axis), respectively. The horizontal display ratio is enlarged to nH and V, n being the integer multiplying factor. In other words, several LCD panels, illumination optics, and circuits are utilized to horizontally enlarge the virtual image from previous 4:3 or 16:9 display ratios to more than twice 8:3 or 32:9 display ratios. In order to create multichannel virtual images simultaneously, two virtual images are located at a single combiner-type lens, as shown in Fig. 3.

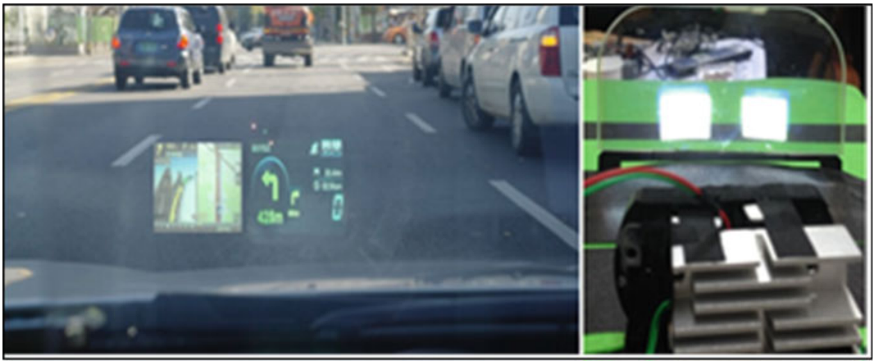


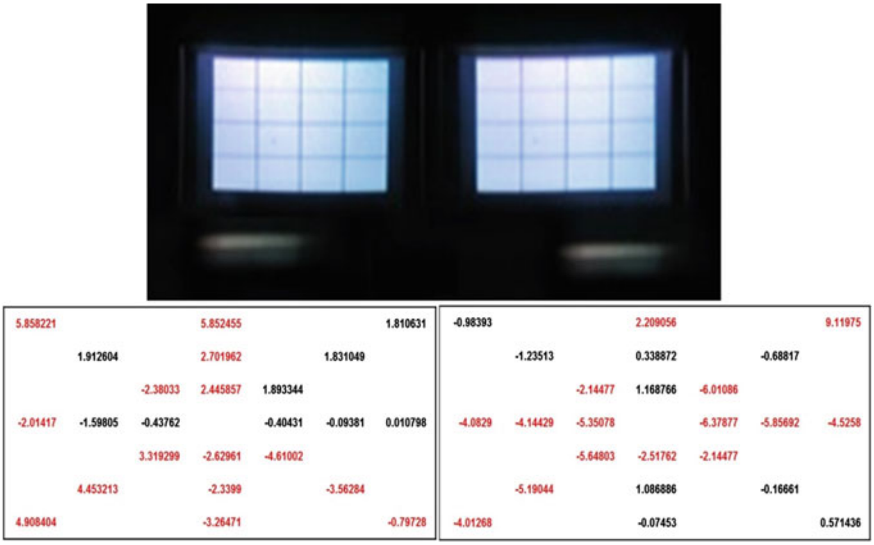
Fig. 4 A back light unit and system design of the HUD using dual panels and LEDs for a multi windows

In order to overcome the shortcoming of not being able to enlarge the virtual image display due to the spatial limits in the vertical direction of the windshield, a multichannel virtual image display that enlarges the display in the horizontal direction is configured. The picture on the left in Fig. 4 is an example of an activated HUD via wireless, displaying a navigation program and mini navigation information from a smartphone. The picture on the right shows white images created by only the light source without an image signal in order to check the uniformity of the brightness between the two displays. The brightness uniformities of the two displays are 83.4 and 82.2%, and considering the structural characteristics of the LCD panels, this is an excellent performance. The size of the two virtual images is 8 inches each and their brightness is designed to vary according to the external level of illumination.

4 Optical System Evaluation

The overall design was optimized to provide the optical performance required. The purpose of the optimization was to fulfill all limiting conditions provided and to secure the optimal performance possible. Distortions at all zoom positions and the modulation transfer function performances were evaluated during the optimization. Examples of performance evaluation elements are the input image and virtual size. Among them, the evaluation of the level of distortion of the two most important image displays was analyzed, shown in Fig. 5.

The image distortion is shown as a pincushion at the edges of the display and it was less than 5% of the standard distortion hence showing a satisfactory result.



5.858221			5.852455			1.810631
	1.912604		2.701962		1.831049	
		-2.38033	2.445857	1.893344		
-2.01417	-1.59805	-0.43762		-0.40431	-0.09381	0.010798
		3.319299	-2.62961	-4.61002		
	4.453213		-2.3399		-3.56284	
4.908404			-3.26471			-0.79728

-0.98393			2.209056			9.11975
	-1.23513		0.338872		-0.68817	
		-2.14477	1.168766	-6.01086		
-4.0829	-4.14429	-5.35078		-6.37877	-5.85692	-4.5258
		-5.64803	-2.51762	-2.14477		
	-5.19044		1.086886		-0.16661	
-4.01268			-0.07453			0.571436

Fig. 5 A distortion analysis of dual projected image in the HUD using the combiner lens

5 Conclusion

The current head-up display has limits on the size of the virtual image as it is a built-in-type with antireflecting coating inserted in the windshield of the automobiles. It also has restrictions based on the temperature of the system and in expressing colors. In this study, a multichannel window-type head-up display is configured by applying multiple LCD panels to enlarge the virtual image horizontally. Micro LCD panels and environment-friendly LED lights were used and diverse images were clearly provided to the driver through an optical design technology [11]. The optical structure can consist of either more than two LCD panels, multiple micro-electro-mechanical-systems-type digital mirror displays, or multiple organic light-emitting diode panels.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF 2015R1C1A1A01053249).

References

1. Gish KW, Staplin L (1995) Human factors aspects of using head up displays in automobiles. U.S. Department of Transportation National Highway Traffic Safety Administration
2. Cotton SDO (1996) Colour, colour spaces and the human visual system. School of Computer Science, University of Birmingham, Technical Report, May 1996

3. Okabayashi S, Fukano M, Daidoji S et al (1989) Development of practical heads-up display for production vehicle application. SAE Technical paper 890559, pp 69–71, Feb 1989
4. Weihrauch M, Meloeny G, Goesh TC (1989) The first head-up display introduce by general motors. SAE Technical paper 890288, pp 55–57, Feb 1989
5. Milanović V, Kasturi A et al (2015) High brightness MEMS mirror based head-up display (HUD) modules with wireless data streaming capability. In: SPIE conference on MOEMS
6. Displaybank's LED Division 2010. Analysis of LED and LED applied product's radiation technology, pp 17, May 2010
7. Hunter Lab (2007) CIE 1976 L * a * b* color scale, pp 1–4
8. Shin S, Chi Y et al (2011) Portable virtual display system design with an eye box motion for motor vehicles. In: Proceedings of SPIE 8167, optical design and engineering, vol 4, pp 176–181, Sep 2011
9. Melzer JE, Moffitt KW (1997) Fundamentals of HMD optics. Head-mounted displays: designing for the user. McGraw-Hill, New York, pp 83–90
10. U.S. Department of Defense (1989) Human engineering design criteria for military systems and equipment, and facilities. MIL-STD-1472D, 1989
11. Abu-Ageel N, Aslam D (2014) Laser-driven visible solid-state light source for etendue-limited applications. J Disp Technol 700–703

Yongseok Chi received the M.S. and Ph.D. degrees in electronics and electrical engineering from Dankook University in 2008 2014, respectively. From 1998 to 2000 he was a research engineer at Elevator Research Lab. in LG industrial company. From 2000 to 2004, he was a research engineer at PDP Research Lab in Samsung SDI Co., Ltd and from 2000 to 2012 he was a chief research engineer at Materials and Devices Research Lab. in LG Electronics. Currently he is an Assistant Professor at Dongseo University in Korea. His research interests are display and optical systems.

Real Time Driver Anger Detection

Afizan Azman, Kirbana Jai Raman, Imran Artwel Junior Mhlanga, Siti Zainab Ibrahim, Sumendra Yogarayan, Mohd Fikri Azli Abdullah, Siti Fatimah Abdul Razak, Anang Hudaya Muhamad Amin and Kalaiarasi Sonai Muthu

Abstract The field of artificial intelligence has seen an increasing number of researches being done related to facial expression recognition. Different methods have been proposed with some of them yielding good results and some performing poorly. Apart from that, anger plays a pivotal role in road accidents since road rage is stated to be one of the contributing factors to road accidents. In order to cater road rage and considering it being harmful to drivers and passengers, this paper proposes a real time driver anger detection. The project classifies human facial expressions, mainly anger expression in real time from a live video in order to warn the driver and eventually road accidents can be reduced.

A. Azman (✉) · K. J. Raman · I. A. J. Mhlanga · S. Z. Ibrahim · S. Yogarayan ·
M. F. A. Abdullah · S. F. A. Razak · A. H. M. Amin · K. S. Muthu
Faculty of Information Science and Information (FIST), Multimedia University (MMU),
Malacca, Melaka, Malaysia
e-mail: afizan.azman@mmu.edu.my

K. J. Raman
e-mail: jpk_kirbz@hotmail.com

I. A. J. Mhlanga
e-mail: imranartweljunior@gmail.com

S. Z. Ibrahim
e-mail: sitizainab.ibrahim@mmu.edu.my

S. Yogarayan
e-mail: mastersumen@gmail.com

M. F. A. Abdullah
e-mail: mfikriazli.abdullah@mmu.edu.my

S. F. A. Razak
e-mail: fatimah.razak@mmu.edu.my.edu.my

A. H. M. Amin
e-mail: anang.amin@mmu.edu.my

K. S. Muthu
e-mail: kalaiarasi@mmu.edu.my

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_17

Keywords Classifier · SVM · Haar cascade classifier

1 Introduction

Humans communicate with each other using verbal as well as non-verbal means. Most of the time non-verbal cues are used to complement the words. Amongst the non-verbal cues that humans use to express their thoughts and feelings are facial expressions. Ekman concluded from his study that facial expressions are universal and innate [1]. This means that even though there might be some variations on how different people express emotions, there are some common features similar on images showing the same emotion. Nowadays, it is believed that for humans and computers to interact effectively, there is need for the humans and computers to interact naturally. This of course is a difficult task due to variations in how people express their emotions.

A research conducted by The Star Online states that Malaysian drivers holds the highest record in accidents caused by road rage [2]. As described earlier, this paper proposes a real time driver anger detection system using webcam, which is focused for driving environment. The system focuses firstly on detecting the human face from the video stream, capturing the face and classifying the human emotions into anger or not angry based on the features from the captured face. If anger is result of the classification is anger, then an alert sound is activated to alert the driver. The proposed work is described along with the results and discussion. In Fig. 1 shows the high-anger drivers according to state in Malaysia.

2 Findings

Malaysia has seen a lot road accident due to driver carelessness. According to The Star, 18% of Malaysian drivers have high-anger issues thus posing a big threat to road safety [2]. Many claim that lack of patience often result in drivers venting while driving with Road Safety Department director-general Datuk Dr. Tam Weng Wah describing drivers as being very less courteous [3]. Dr. Tam also pointed out that the belief by many divers that they should always be first often leads to wrong judgements thus leading to accidents. While the causes for anger driving span from personal to environmental issues like police presence, slow driving by other motorists and traffic obstruction. Results of a study published by The Star News shows that Malaysian drivers have a high rate of traffic discomfort due to reasons like police presence, slow driving and illegal driving as compared to other countries as shown on Table 1 [3].

Furthermore, another study published by The Star News again shows alarming levels of anger driving in different states of Malaysia with Terengganu, Malacca and Kuala Lumpur recording 23.2%, 22.2% and 21.8% respectively as shown on Table 1 [2].

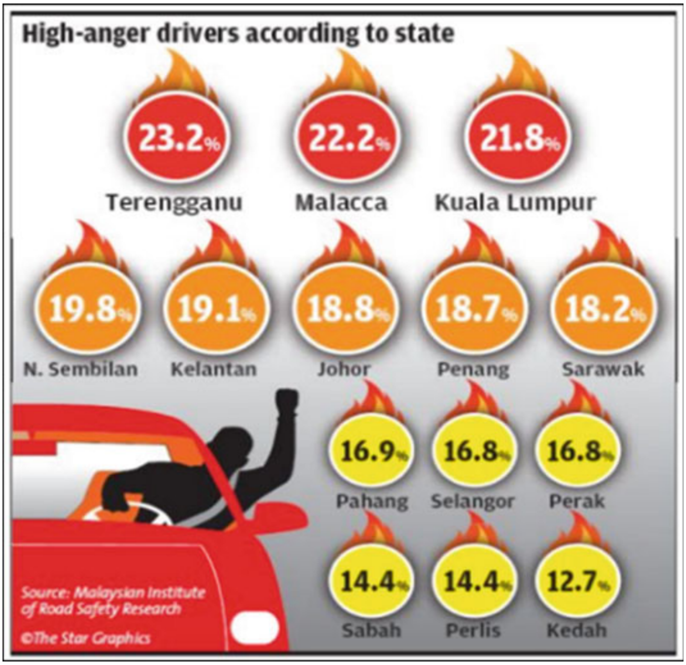


Fig. 1 High-anger drivers according to state [2]

Table 1 International comparison of anger [2]

Country	Hostile gestures	Illegal driving	Police presence	Slow driving	Traffic obstruction	Discourtesy
USA	3.2	2.7	3.0	3.2	3.3	3.9
Britain	2.3	2.3	1.4	2.0	2.0	2.7
Australia	2.8	2.6	1.9	2.4	2.3	3.1
N.Zealand	2.7	3.3	1.9	2.8	2.7	3.5
Spain	2.9	3.5	2.0	2.3	2.8	3.7
Turkey	3.4	3.5	2.2	2.9	3.1	3.6
Malaysia	6.3	7.0	4.3	6.1	9.2	10.0

With these results at hand, it is however safe to conclude that anger driving is a major concern in Malaysia and it has to be addressed proactively. Thus anger detection system is a step in trying to address this concern. For it to be effective, it requires accurate and timely classification and detection of anger which means effective and accurate algorithm need to be designed.

Table 2 Summary of previous facial expression research

Author(s)	Title	Year	Method
David Matsumoto and Hyi Sung Hwang	Reading facial expressions of emotion [5]	2011	Microexpression
Zainal Abidin and Harjoko	A neural network based facial expression recognition using fisherface [6]	2012	Neural network + fisherface
Jyoti Kumari, R Rajesh, K M Pooja	Facial expression recognition: a survey [7]	2015	LBP, LDP, HOG
Szwoch and Pieniazek	Facial emotion recognition using depth data [8]	2015	Depth image
André Teixeira Lopesa, Edilsonde Aguiarb, Alberto F. De Souzaa, Thiago Oliveira-Santosa	Facial expression recognition with convolutional neural networks [9]	2017	CNN

2.1 Facial Expression

The greater part of human communication is by non-verbal means. Body movements, gestures are often used to complement verbal communication. Moreover, most of the time, body movements are usually a reflection of a person's emotion and facial expression is one of the many ways to show emotions. There are seven universal emotions namely happiness, sadness, fear, disgust, anger and surprise. Ekman pointed out in his research that facial expressions are universal with an exception of some cultures like the Japanese where people hide their emotions in the presence of high authority [1]. However, this project will focus mainly on anger.

Anger is one of the universal facial expression. Regardless of tribe, race or culture the expression always be the same. Generally, anger is commonly expressed in situation a person assess as being unpleasant, frustrating or irritating. According to a research, there are common characteristics found in almost every angry person, which are [4]:

- Eyebrows slanting inwards and lowered and at the same time squeezed
- Straight and tight eyelids due to lowered brows
- Puckered lips, mouth closed as straight tight lips or open in a yelling gesture
- Forward thrust of the jaw

In Table 2 the summary of previous research of facial expression using different methods is shown. There are various approaches used in the study and that is inclusive of microexpression, neural network, fisherfaces, local binary pattern, local derivative pattern, histogram of oriented gradients, depth image and convolutional neural networks.

2.2 Haar Cascades

A cascade is a combination of features (Haar features). The Haar features represent a rectangular pattern of data. These features can be edge features representing the edge of an object, line features or centre-surround features [10]. The use of haar feature classifier is an effective method of detecting objects which was initially suggested by Paul and Viola in a research paper, "Rapid Object Detection using a Boosted Cascade of Simple Features." This cascade is trained with a lot of positive and negative images and in the end it is used to detect objects in other images. In as far as detecting a face for example, these features are combined to form a classifier. This classifier is considered a weak classifier as its accuracy is barely above 50%. When these weak classifiers are combined in a particular order, they form a strong classifier which is a cascade of classifiers which will be able to detect the features it was trained for. This is achieved through a process of AdaBoost suggested by Viola and Jones. AdaBoost tries out multiple weak classifiers over a series of rounds and selected the best weak classifier in each round. At the end the selected best weak classifiers are the combined to form a strong classifier which has a higher accuracy than a single weak classifier [11].

2.3 Support Vector Machine (SVM)

This is a supervised learning model that is allied with different algorithms that analyse data for classification as well as regression. It can perform linear as well as non-linear classification. This project will however make use of linear classification. In used in a linear classification, an SVM model plots the points on a group which corresponds to the two classes of data. After plotting a graph of points, it computes hyperplane which is the maximum line that separates the two classes.

The graph in Fig. 2 shows two classes of features, circles and squares. The two classes are then separated by a red line (hyperplane). The main goal of SVM is estimated the optimum hyperplane such that the distance between the dotted lines is large. When a new feature is introduced, its classes is determined by which class of features is it close to. If it is close to the squares class, then the classifier will predict it as to be belonging to the square class. However, if the new feature lies on the hyperplane, the classifier will neither predict it as square nor circle. The predicted label will be 0 (zero).

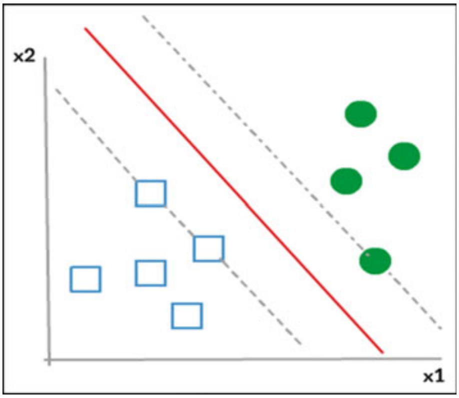


Fig. 2 Support vector machine

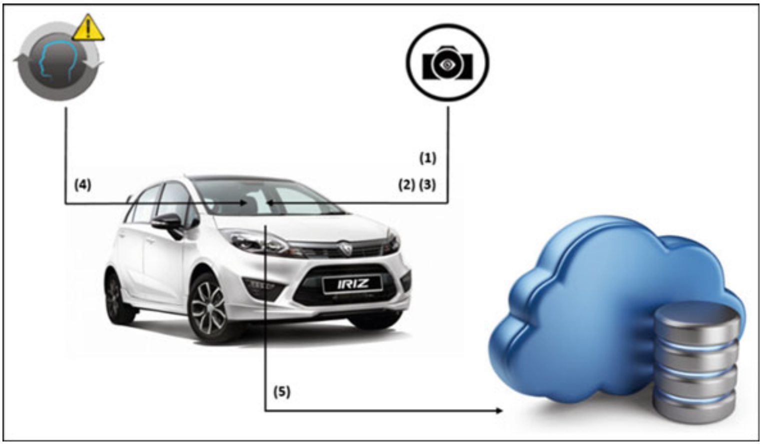


Fig. 3 Process of driver anger detection system

3 Proposed System

The deliverable of the implementation phase will be a real time system that detects anger on the driver's face. There will be a webcam that will be placed in front of the driver. This webcam will be used to stream a live video. From this video, video frames will be captured and from these frames face region will be detected. The face region will be cropped and converted to grayscale to simplify the feature extraction process. Images collected during the planning phase will be used for training the classifier and the captured driver's images will be tested against the classifier and if anger expression is detected then an alert sound will be triggered. The entire flow of the system is as shown in Fig. 3.

In order to develop the system, OpenCV 3.0 library was used. This is a powerful computer vision library under BSD licence which was designed mainly for real time applications. It contains over 500 algorithms some of which are used for detection as well as recognition. Many organizations use OpenCV for different purposes ranging from facial recognition, car tracking to as far as augmented reality. This is due to the fact that it comes with some powerful algorithms like artificial neural networks, k-nearest neighbour algorithm. OpenCV works well with Python programming language which also is a simple, easy and powerful language. However, in this project OpenCV will be used for face detection as well as classification.

4 Result and Discussion

4.1 Result

The system was tested on several participants. The participants were asked to express anger and any other expression. The following snapshots were taken while the system was running. Detection of anger is shown in Fig. 4 while a non-angry face detection is as shown in Fig. 5.

However, there were times when the model misclassified some images. This could have been because of sitting position relative to the camera and also lack of variety of images in the anger class. A misclassified face as anger is as shown in Fig. 6.

4.2 Validation

After the instantiating the support vector machine classifier, the classifier had to be validated in order to calculate the confidence of the classifier. In order to perform this

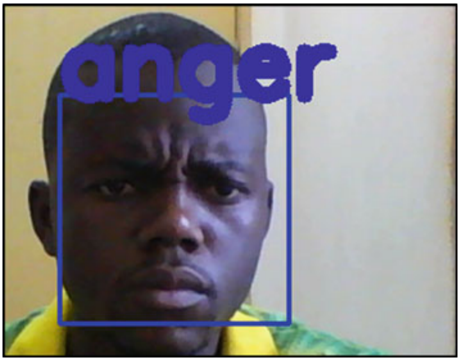


Fig. 4 Anger detection

validation a technique called cross validation was implemented. JAFFE database as shown in Fig. 7 is used for training purposes.

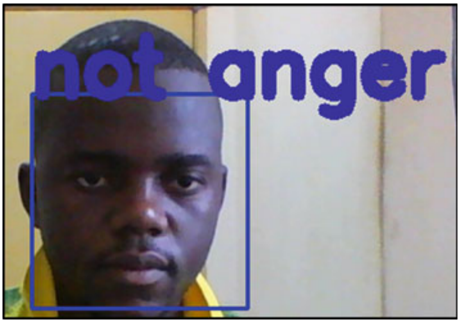
This is a model evaluation technique used in predictive analysis to estimate the accuracy of the trained model when used in practical situations. In this project, K-Fold cross validation method will be used to evaluate the model as shown in the function below:

```
_cross_validation (clf, X, y, k)
cv = KFold(len(y), k, shuffle=True, random_state=0
scores = cross_val_score(clf, X, y, cv=cv)
```

The function will take in the instantiated SVM classifier together with the train data, train labels and a constant value k. The data set will be divided into k subsets and the cross validation holdout method is repeated k times. Each time a k subset is used as the testing set and the other $k - 1$ subset is used as the training set. This is to say that the classifier is tested to check if it can predict correctly the data which is in subset k after having been trained using the data in k − 1 subsets. At the end the average error is computed for all the k trials. The mean score will be shown which will show just how much the model is confident (in terms of percentage) in predicting emotions on images it has not seen before.

In Fig. 8 below, shows the results of cross validation done using 5-folds. The mean score indicates that the model is about 97% confident of predicting new images accurately.

Fig. 5 Not angry detection



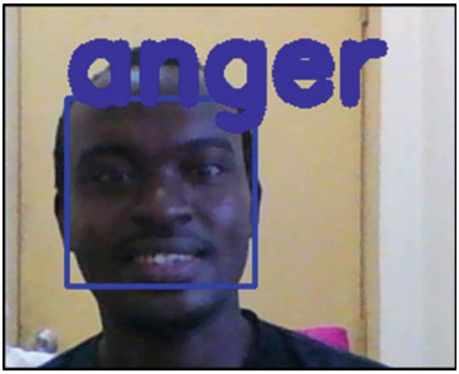


Fig. 6 Misclassified anger detection

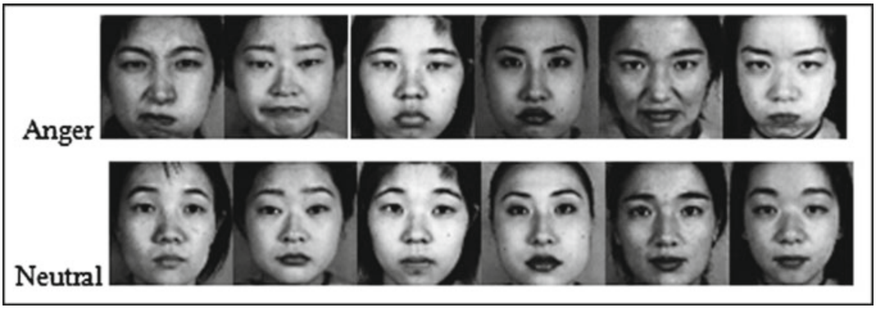
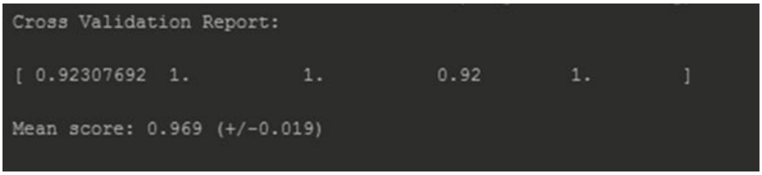


Fig. 7 JAFEE database images [12]



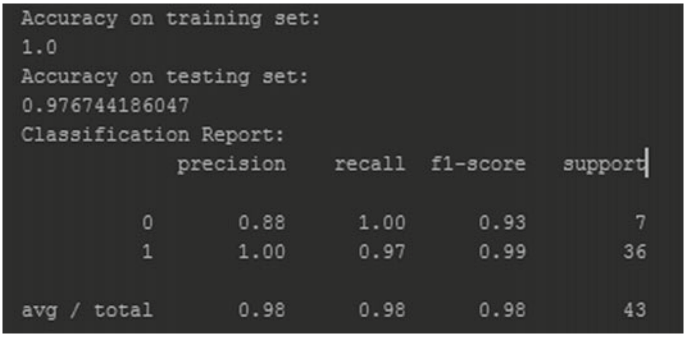
```
Cross Validation Report:

[ 0.92307692  1.          1.          0.92        1.        ]

Mean score: 0.969 (+/-0.019)
```

Fig. 8 Cross validation results

4.3 *Training*

Training will be done using JAFEE database. The database is a set, which contains 213 images showing all the 7 expressions posed by 10 Japanese females. All the image samples will be normalized to 70 × 70 pixels. Below are some of the images from the database.

This will be the last step of the system before testing. The cross-validated SVM classifier will be tested using the images in the X_test list. Accuracy will be measured by comparing the output of the prediction with the correct label of the image used



```
Accuracy on training set:
1.0
Accuracy on testing set:
0.976744186047
Classification Report:
             precision    recall  f1-score   support

          0       0.88      1.00      0.93         7
          1       1.00      0.97      0.99        36

avg / total       0.98      0.98      0.98        43
```

Fig. 9 SVM classification report

for testing. SVM built in fit function will be used to train the classifier and another function, predict will then be called to make a prediction on the data.

```
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(metrics.classification_report(y_test, y_pred))
```

In Fig. 9, the precision values shown represents the ratio of the expressions the classifier was able to recall correctly to the number of the expressions the classifier recalled which is a mix of correct and wrong recalls. The recall values presents the ratio of the number of correctly recalled expressions to a number of all the correct expressions relative to the images. In the case of the Cohn Kanade (CK) database where the images of anger very much less than the neutral images, using precision is the better way to evaluate the classifier. The f1-score or also known as f-measure represents the harmonic mean between recall and precision. This is because an algorithm can have high recall and less precision [13, 14].

5 Conclusion and Future Works

In this paper an approach was presented for detecting facial expression mainly anger in real time using Support Vector Machine together with Viola Jones Haar feature algorithm. For face detection, a Haar cascade file was used to detect the frontal face. After the face was detected, an RGB (Red Blue Green) image was captured. Due to the complexity of feature extraction on an RGB image, the captured image was converted to grayscale and was resized to set standard size of 70 × 70 pixels to match also the size of the images in the training set, which will be used to build the classifier. Then an SVM model was instantiated and was cross validated using K-Fold cross validation approach. Finally, prediction was done and an alert sound was activated once anger is detected for about 3 s.

In the future, a convolution neural network will be used for the same problem notably the one developed by Google called Inception. This network is trained for ImageNet using over 1.2 million images. It can be used to differentiate between one thousand different classes. It can be used to detect almost anything by just retraining the last layer of the network. The project intend to implement facial expression recognition in areas were public engage like banks. This will help the bank owners to track the satisfaction of their customers. It is also believed that if customized this system can also be used for event monitoring. This will be very helpful in big events were so many things are going on so that the organizers can also recognize the valence of the people present.

Acknowledgements This project is externally supported by Multimedia University (MMU), Melaka, Malaysia and carried out research and development under the Connected Car Services team.

References

1. Ekman P (1994) Strong evidence for universals in facial expressions. Psychol Bull https://doi.org/10.1037/0033-2909.115.2.268
2. The Star Online (n.d.) Find out root of aggressive driving style, government urged. http://www.thestar.com.my/news/nation/2013/07/15/anger-management-needed-find-out-root-of-aggressive-driving-style-govt-urged/
3. Carsifu (2016) The typical Malaysian driver is selfish and delusional. http://www.carsifu.my/news/The-typical-Malaysian-driver-is-selfish-and-delusi
4. Bartlett MS, Littlewort G, Fasel I, Movellan JR (2003) Real time face detection and facial expression recognition: development and applications to human computer interaction. In: 2003 conference on computer vision and pattern recognition workshop, vol 5, pp 53–53
5. Matsumoto D (2011) Reading facial expression of emotion, psychological science agenda
6. Abidin Z, Harjoko A (2012) A neural net-work based facial expression recognition using fisher-face. Int J Comput Appl 59(3):30–34
7. Jyoti K, Rajesh R, Pooja KM (2015) Facial expression recognition: a survey. In: 2nd international symposium on computer vision and the internet (VisionNet'15)
8. Szwoch M, Pieniążek P (2015) Facial emotion recognition using depth data. In: 2015 8th International conference on human system interaction (HSI), Warsaw, pp 271–277
9. Lopes AT, de Aguiar E, De Souza A, Santos TO (2016) Facial expression recognition with convolutional neural network
10. Stan Z, Anil K (2011) The handbook of facial recognition. Springer, New York
11. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features
12. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J Coding facial expressions with gabor wavelets. In: Third IEEE international conference on automatic face and gesture recognition
13. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the fourth IEEE international conference on automatic face and gesture recognition (FG'00), Grenoble, France, pp 46–53
14. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete expression dataset for action unit and emotion-specified expression. In: Proceedings of the third international workshop on CVPR for human communicative behavior analysis (CVPR4HB 2010), San Francisco, USA, pp 94–101

Debug Output Features for OpenGL SC 2.0 Safety Critical Profile

Woosuk Shin and Nakhoon Baek

Abstract With the modern 3D graphics libraries, the support for the error reports and other debugging features are much important. For the famous OpenGL (Open Graphics Library) family, they introduce the KHR debug extension. In this paper, we present the details of adding full features of KHR debug extension to the OpenGL SC (Safety Critical profile) 2.0. This enhancement to the OpenGL SC 2.0 engine shows more up-to-date debugging features with traditional graphics drivers.

Keywords OpenGL · Safety critical profile · Debug output extension

1 Introduction

Currently, *OpenGL* family is one of the most widely used three-dimensional graphics library [1, 2]. This library is used on various platforms including personal computers, workstations, mainframes, tablets, smartphones, and others. Many commercial and non-commercial implementations are available. For three-dimensional graphics output, the single board computing devices typically use *OpenGL ES* (Open Graphics Library for Embedded Systems) [2]. *Khronos Group*, the de facto standard organization, consistently manages all the standard specifications and various extensions.

From the programming point of view, the emphasis on the debugging supports are one of the most important trends. In contrast, OpenGL graphics library was originally designed in 1990s, when there is no sufficient requirements on the debugging

W. Shin · N. Baek (✉)
School of Computer Science and Engineering, Kyungpook National University,
Daegu 41566, Republic of Korea
e-mail: oceancru@gmail.com

N. Baek
Software Technology Research Center, Kyungpook National University,
Daegu 41566, Republic of Korea

N. Baek
dassomey.com Inc, Daegu 41566, Republic of Korea

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_18

functionalities [3]. Thus, many existing drivers and implementations of OpenGL provides only the classical error checking feature: When an error occurred, OpenGL system set the internal error flag, and it is the duty of application programmers to periodically check the error flag and process its error handling.

OpenGL SC (OpenGL for Safety Critical) is conceptually a safety critical variation of the famous OpenGL standard. This graphics API library is designed to meet the needs of safety critical markets for avionics, industrial, military, medical and automotive applications. In the case of safety-critical markets, OpenGL SC plays the major role for the graphical interfaces. The need for this 3D graphics API is rapidly increasing with the growth of the safety-critical market [4, 5]. For the medical and automotive applications, consumer electronics markets start to strongly need this standard [6].

In the year of 2015, the Khronos Group, the fundamental standard management body of the OpenGL family, established the new OpenGL SC 2.0 specification [7]. It originally aims to a safety critical subset of OpenGL ES 2.0 [8]. The Khronos Group is also developing cross-API guidelines to aid in the development of open technology standards for safety critical systems.

For the safety critical profiles, they also emphasize the security-critical features. To fully support these safety and security features, the debug output extensions are highly recommended. In this paper, we show our design and implementation of the debug output extension for OpenGL SC 2.0 library. Our implementation shows that it works well and helpful for the safety-critical supports. Details are followed in the following sections.

2 Design Analysis

For more convenient error processing and debugging features, the new *KHR debug extension* [9] to the OpenGL is introduced in 2012. With this debugging extension, the OpenGL system generates more detailed internal error messages and automatically calls the pre-registered callback function when an error occurred. This new extension is actually a big improvement to the debugging of the OpenGL application programs. From OpenGL version 4.3 [10], this debugging extension is now the core feature of the OpenGL standard specification.

With KHR_debug extension, the OpenGL system automatically calls the pre-registered callback function, for every internal errors. We can use the `glDebugMessageCallback` function to pre-register the user-specifiable callback function. Additionally, application programmers can generate his/her own messages to the error stream, even without any OpenGL errors. We can use `glDebugMessageInsert` function for this purpose. The followings are the details of those functions.

```
void glDebugMessageCallback(DEBUGPROC callback, const void* userParam);
```

specifies a callback function to receive debugging messages from the OpenGL. Each time a debug message is generated the debug callback function will be invoked with source, type, id, and severity associated with the message.

```
void glDebugMessageInsert(enum source, enum type, uint id, enum severity,
    sizei length, const char* buf);
```

injects an application-supplied message into the debug message queue.

When the OpenGL system works in the server-client model, the OpenGL functions may be called with remote calls. In this case, since the physical memory regions are isolated to each side, the OpenGL system cannot call the pre-registered callback function. For those situations, the OpenGL system generates the debugging message to its internal message log area. Application programmers can access the internal message log area and get the log messages using the following function:

```
uint GetDebugMessageLog(uint count, sizei bufSize, enum* sources, enum*
    types, uint* ids, enum* severities, sizei* lengths, char* messageLog);
```

retrieves messages from the debug message log.

We applied the KHR debug extension to the new OpenGL SC 2.0 standard specification. To support the debug extension, we first need API function implementation. To support full features, actually we need internal engine modifications, to support debug features. In our case, we succeeded to add all the features to the OpenGL SC 2.0 engine internals.

3 Implementation Results

To test the KHR debug extension features, we make a set of example programs. As a typical example, one of our example source code contains the following code segment:

```
GLint loc_color2 = glGetAttribLocation(program_710, "unspecified");
glEnableVertexAttribArray(loc_color2);
```

Since our shader programs do not contain any variable of the name “unspecified”, these code segments generate a warning and an error, according to the OpenGL SC 2.0 standard specification.

In our OpenGL SC 2.0 engine implementation, each API function will check any warning conditions and error conditions. For the warning case of glGetAttribLocation function, our engine contains the following code segment:

```
if (iAnswer == -1) { // warning condition found
  int level = WARNING;
  int id_num = 409;
  makeDebugMesg ( level, id,
    "%s: program(=%u) does not have an attribute named [%s]",
    pszFunc, program, name );
}
```

where makeDebugMesg( ) is the internal debug output function. This will generate the corresponding OpenGL SC warning debug message.

Additionally, our OpenGL SC 2.0 engine also contains error check as follows:

```
if (index == static_cast<GLuint>(-1)) { // error condition
  makeErrorMesg( GL_INVALID_VALUE, uiFunc,
    "%s: index(=%d) is an INVALID location", pszFunc, -1 );
    goto done;
}
```

Finally, these warning and error checks generate the debug output messages on the screen as follows:

```
OpenGL WARNING: UNKNOWN [API,OTHER,LOW,id=409] glGetAttribLocation:
    program(=33554433) does not have an attribute named [in_unspecified]
OpenGL ERROR: INVALID_VALUE [API,ERROR,HIGH,id=213] glEnableVertex-
    AttribArray: index(=-1) is an INVALID location
```

As shown here, these debug output messages contain the API kind, severity level, and identification numbers, as the details of the debug messages. All these details are not required for the basic-level OpenGL SC engines, while those should be implemented for the KHR debug extension support.

4 Conclusion

In these days, the debugging features become more important to the development process. In this paper, we aimed to add the KHR debug extension to the newly released OpenGL SC 2.0 graphics library. We designed to support all the debug output features in the existing OpenGL SC 2.0 engine. The final check programs show that our OpenGL SC 2.0 engine now fully support KHR debug extensions. This is the first literature report of the debug extension to the OpenGL Security Critical profile, at least to the best of our knowledge.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant 2016R1D1A3B03935488).

This study was also supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005).

References

1. Khronos Group (2015) OpenGL 4.5 core profile
2. Khronos Group (2012) OpenGL ES version 3.0
3. Segal M, Akeley K (1994) The OpenGL graphics system: a specification (version 1.0), silicon grpahics
4. Cole P (2005) OpenGL ES SC—open standard embedded graphics API for safety critical applications. In: 24th digital avionics systems conference
5. Snyder M (2005) Solving the embedded OpenGL puzzle—making standards, tools, and APIs work together in highly embedded and safety critical environments. In: 24th digital avionics systems conference
6. Baek N, Baeck GJ (2010) Design of OpenGL SC emulation library over the desktop OpenGL 1.3. In: 29th digital avionics systems conference
7. Fabius A, Viggers S (2016) OpenGL SC, version 2.0.0 full specification
8. Munshi A, Leech J (2010) OpenGL ES common profile specification, version 2.0.25 (Full specification)
9. Riccio C (2012) KHR_debug extension. Khronos Group
10. Khronos Group (2012) OpenGL 4.3 core specification

Eigennose: Assessing Nose-Based Principal Component Analysis for Achieving Access Control with Occluded Faces

Zibusiso Bhango and Dustin van der Haar

Abstract State-of-the-art face recognition systems exist today with varying performances. However, many suffer from multiple occlusions that threaten their performance. The common causes of these occlusions are hats, scarves and, sunglasses. Usually, when occlusions are present, the nose features are available. Surprisingly, not much research has been focused on nose biometrics. Research has shown that the nasal area provides robust, discriminant features that can be used to positively authenticate a user. In our system, we attempt to authenticate a user using only their nose. Eigennose algorithm, which is an extension of the eigenface algorithm is developed to find the discriminant nasal features of individuals with Euclidean distance used for matching. The system is then compared with machine learning algorithms such as Support Vector Machines and k-Nearest Neighbor to find better-performing methods. Our experiment did not achieve very good performance.

Keywords Nose recognition · Face occlusions · Access control · Eigennose

1 Introduction

Since the mid-20th century, face recognition systems have played an integral part in identifying, authenticating and monitoring people. Face recognition systems have since emerged as the natural solution to identifying people being authenticated or

Z. Bhango · D. van der Haar (✉)
Academy of Computer Science and Software Engineering, University of Johannesburg,
Cnr University Road and Kingsway Avenue, APK Campus, Johannesburg 2006, South Africa
e-mail: dvanderhaar@uj.ac.za

Z. Bhango
e-mail: zbhango@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_19

monitored. Today, state-of-the-art face recognition systems exist with good performance under varying luminance, facial expressions, lighting etc.

However, many face recognition systems underperform when partial face occlusions caused by objects such as sunglasses and scarves are present [13] . Many face recognition systems need users to look directly into the camera for a specific amount of time for them to acquire all the necessary features to authenticate the users. Face recognition systems can struggle with identifying faces with significant differences in expressions such as smiling, laughing or frowning.

A potential solution to these problems is a biometric system that works optimally with multiple face occlusions. In many scenarios where face occlusions occur, the nose is visible because humans use their nose to breathe and hiding one's nose in public is perceived as suspicious behavior. This research attempts to use this information to build a robust nose biometrics system for access control.

If using the nose for access control is faster, it might provide a performance gain. The nasal area is invariant to facial expressions and weight gain/loss, ensuring a consistent performance over time. If nasal features are unique, then they can be used for access control in a profile view environment with occluded faces, vastly improving on the shortcomings of face recognition systems in such scenarios.

The rest of the paper aims to compare the performance of the eigennose algorithm with linear classifiers against machine learning classifiers. The next sections are divided as follows. In Problem Background, we focus on the problem at hand, which is how to achieve access control with occluded faces. In Related Work, we describe tried methods in literature that tackle similar problems to ours. In Experimental Setup, we describe our proposed approach in detail and explain the methods that we will use to preprocess, extract and classify nasal features. In Results, we provide results on our methods and compare the methods used to find the best performing. In Conclusion, we provide our findings from the research and future work.

2 Problem Background

It has become necessary for commercial industries to deploy biometric systems to enhance their security. Positive authentication before gaining access to business or government assets has become necessary. Biometric systems used to solve these problems include fingerprint recognition, iris recognition and voice recognition, among others. These systems have had different levels of success, each with unique advantages and disadvantages. Iris recognition is unique, discriminative and robust but is very difficult to capture and can easily be occluded by eyelids or sunglasses [6]. Ears have a consistent structure and are easy to capture, but can be easily occluded by wearing hats [1].

Gait recognition is effective for monitoring someone at a distance but many factors are against its level of practicality [8]. The choice of footwear, the nature of clothing, affliction to the legs and the walking surface can heavily alter the way one walks [10].

Face recognition has been the most successful and widely used biometrics system in surveillance environments and is second only to fingerprint recognition [3]. Face recognition is cheaper, its features are more flexible to acquire and many tried and tested algorithms exist in literature. However, many face recognition systems suffer from occlusions. When instrumental parts of the face are not available for feature extraction, these algorithms struggle to classify and match correctly, leading to an increase in false negative rate and user inconvenience.

Eidenberger used Kalmanfaces to develop a face recognition system that performs well under varying illumination conditions to cater for poor lighting conditions [2]. Khorsheed and Yurtkan used local binary patterns to authenticate individuals under varying facial expressions based on anger, disgust, fear, happiness, sadness, and surprise [4]. Oscos, Khoshgoftaar, and Wald developed a rotation invariant face recognition system that authenticates individuals with in plane and in depth rotations [9].

Face occlusions have troubled face recognition systems. This is because when important features are not provided for authentication, the system's performance becomes an issue no matter how good the algorithms are. Nose biometrics is a potential solution to this problem. In most cases where multiple face occlusions are present, the nose is available. Humans use the nose to breathe and hiding your nose in a public area can be regarded as suspicious behavior.

Studies have shown that when facial expressions change, or people gain or lose weight, the nasal features are rarely affected. The nose features characteristics include discriminability, robustness, and ease of extraction, and the nose is arguably the part of the face containing the most significant information [13]. These are great characteristics for a biometric system for access control and can prove to be a great solution to be used together with face recognition systems when face features are occluded.

3 Related Work

Turk and Pentland developed a face recognition system that treats a face as a 2-D recognition problem, using the assumption that faces are always upright in an image [11]. The system used eigenfaces, derived from eigenvectors, which do not correspond to face features such as nose, ears or eyes. These eigenvectors can be thought of as principal components of the distribution of faces. Each eigenface represents a difference in variation among the face images used.

Their method was very successful in locating and recognizing face images and classifying them. However, because they used a principal component analysis that does not recognize geometric features independently but instead uses spatial or texture characteristics, it may struggle for performance when face images look alike as a whole but features are different individually.

Moorhouse et al. used photometric stereo images to propose two-dimensional features for nose recognition [6]. They used Fourier descriptors to capture the ridge shape from the nasal tip and used geometric ratios of these features to classify noses.

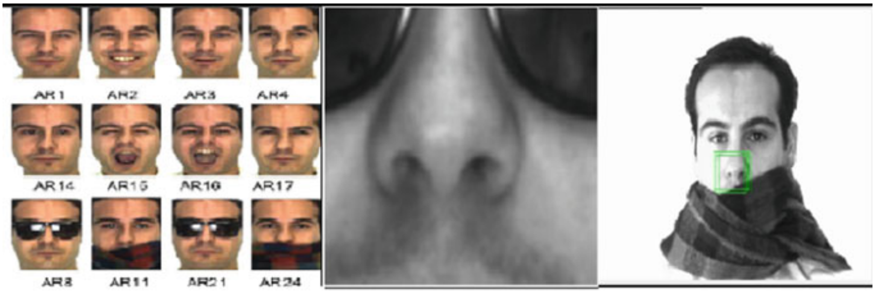


Fig. 1 AR database used for data sampling to test the method (left). A human nose detected from an occluded face image (center). A nose detected using the Viola-Jones object detection algorithm (right)

They classified noses into six classes: Greek, Nubian, Roman, Snub, Turn-up, and Hawk. Although their classification algorithm was good, their recognition rate was 10%. Also, using such special equipment deems the research impractical in most cases because such equipment is expensive to acquire and needs calibration. They also used a private dataset to measure their system performance, which makes it difficult to truly measure its performance.

Zuo et al. developed a face recognition system that attempts to handle occluded faces, using IRF-Eigenfaces [14]. The algorithm differs with the general eigenfaces algorithm in that it first defines an objective function, and then use Iteratively Reweighted Fitting least-squares fitting algorithm to extract feature vector by minimizing the objective function. They used the AR dataset. Their recognition rate was 89%, beating the original eigenface, which had a 37% recognition rate.

These solutions address specific problems, however, the presence of face occlusions will still impact performance. The research aims to provide access control using nasal biometrics with occluded faces (Fig. 1).

4 Experimental Setup

4.1 Data Sampling

For data sampling, we used the AR face database which consists of 20 males and 13 females [5]. Each individual has thirteen images in different lighting and/or facial expression. Face occlusions are present with the use of scarves and sunglasses. 403 images were used for testing the performance of the system.

4.2 Preprocessing

The face image is converted to a grayscale format to remove noise. A histogram equalizer is then used to normalize the image such that the contrast differences follow a uniform distribution. We use this image to detect the nose using Viola-Jones object detection algorithm [12]. The algorithm uses Haar-like features to search the entire image for the nose. To save on computational space and time, each window is converted into an integral image, which is made up of four bits. Not all windows contain the nose features, Adaboost training is used to find the windows best describing the nose features.

We test every window for nose images using Haar-like features. We then use the windows containing the minimum error rates as the windows containing the nose images. After Adaboost training, we use cascading classifiers to find the windows containing nose images for nose classification. Cascading classifiers are split into classes, with each class containing fewer features to check than the next one. Each window is tested for nose images using these classes. Only those images passing the test in each class are sent to other classes for further testing. If an image passes the final class, we have positively identified the nose.

4.3 Feature Extraction

For feature extraction, we used Principal Component Analysis for eigennose. We took multiple nose images of the same individual that are similar in distance (used to capture them) and size and normalized them to get pixel values between 0 and 255. Each image is converted into its 2D array of pixels of size N × N. Each image is then converted into a vector of size $N^2 \times 1$. Characteristic features of each image are calculated by subtracting the features that are common to all images, done using the equation

$$A_i = \frac{1}{M}\sum r_i, \tag{1}$$

where M is the number of vectors and r_i is each vector. To find the characteristic features of each vector C_i, we subtract each vector from A_i.

We convert C_i into a single 2-dimensional array B and multiply it by its transpose B^T. The resulting product will then be used to find the eigenvalues γi and eigenvectors μi. To get the eigennose, we multiply A with each eigenvector μi. The resulting eigennose υ_i is normalized and can be scaled to values between 0 and 255 for visualization. These vectors are then converted into a single 2-dimensional array U, which will be used for classification (Fig. 2).

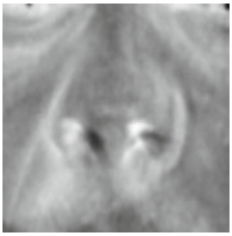


Fig. 2 The resulting eigennose after implementing the eigennose feature extractor

4.4 Classification

A new image for matching is selected and converted into a vector and have its common features subtracted from it. We then multiplied this vector with the transpose of U to get its characteristic features eigenvalues. Every vector in U, that is, U_i is then multiplied with U to get its eigenvalues to be compared with the new image.

For classification, we used the Euclidean Distance. The equation of this classifier is as follows:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}, \qquad (2)$$

Where p and q are vectors being compared against each other for binary classification.

Eigenvalues were used to compare the vectors, with the assumption that a nose image of the same person will have a smaller distance compared to that of a different person. If the distance is less than the heuristic threshold, we have found a match. If not, then the image is not a match.

5 Results

We used Receiver Operating Characteristic (ROC) curve and Detection Error Tradeoff (DET) graph to measure the method's performance. The ROC is used to plot the True Positive Rate versus the False Positive Rate (FPR) to measure the performance of the system when classifying nose images. The Area Under the Curve determines whether the system performs well or not. The DET graph is used to plot the False Positive Rate (FPR) versus the False Negative Rate (FNR). It is used to see the error rates of the binary classifier and to calculate the Equal Error Rate (EER). The EER is when the FPR and the FNR are equal (Fig. 3).

The experiment did not achieve very good performance. The Euclidean distance classifier worked because the distance between nasal features of two or more different individuals tend to be more than that of the same individual. However, there is a high false acceptance rate caused by the small inter-class variation and large intra-class

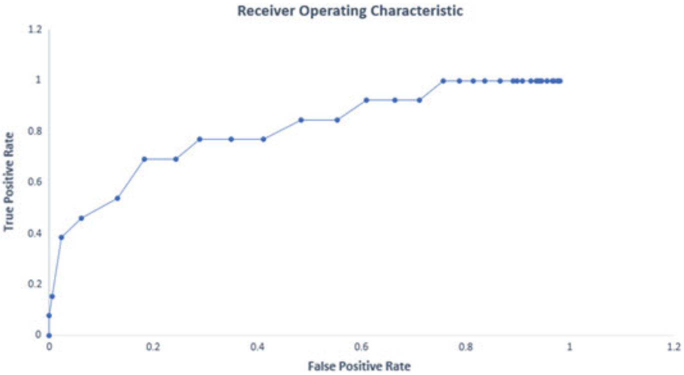


Fig. 3 The receiver operating characteristic curve, plotting false positive rate versus true positive rate

variation of eigennose features. Machine learning approaches used did not perform very good either.

The performance of eigennose and Euclidean distance classifier proves the approach is feasible and can be further improved for better performance. With this performance, we have proven that it is possible to authenticate a user when multiple face occlusions are present, but further research is necessary to improve performance. A nose biometrics system can be used together with a face recognition systems with occluded images to further improve its performance. A nose biometrics system will perform better with occlusions and a face recognition system will perform better when there are no occlusions (Fig. 4).

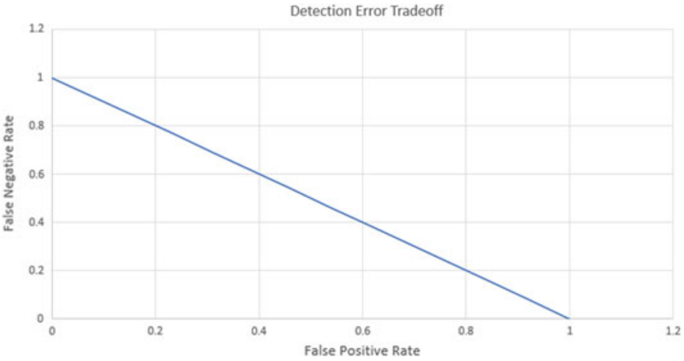


Fig. 4 The DET graph

6 Conclusion

Due to few features available in the nasal area, it is difficult to use independent component analysis algorithms. Using a principal component analysis algorithm such as eigennose is feasible, but not high performing. Eigennose features have a small inter-class and large intra-class variation, making it very difficult to classify nose images and authenticate users. Linear classifiers have high false acceptance rates.

Capturing the nose instead of the whole face makes feature extraction and classification faster, ensuring speedy results and convenience for end-users. However, the performance decreases as well. Due to nasal features characteristics, it is feasible to achieve access control with occluded faces, improving on the shortcomings of the face recognition systems. This research is one of the few on nose biometrics and we believe it will pave the way for further research in this field.

The nose biometrics field is ripe with potential but lacks detailed exploration. 3D nose recognition systems could provide better performance since they add more features to work with, but requires expensive equipment to capture the nose and more time to classify the nose image.

Nose biometrics is deserving of further research because it improves on shortcomings of other biometrics such as the face, ear, iris, and gait. The natural existence of occlusions makes it difficult for these systems to perform, but nose biometrics is not much affected by it. The system performance provides for optimism in the belief that nose biometrics can be a great, robust and efficient form of authentication.

References

1. Abaza A, Ross A, Hebert C, Harrison MAF, Nixon MS (2013) A survey on ear biometrics. ACM Comput Surv 45(2):22:1–22:35. http://0-doi.acm.org.ujlink.uj.ac.za/10.1145/2431211.2431221
2. Eidenberger H (2006) Illumination-invariant face recognition by kalman filtering. Proc ELMAR 2006:69–72
3. Kamgar-Parsi B, Lawson W, Kamgar-Parsi B (2011) Toward development of a face recognition system for watchlist surveillance. IEEE Trans Pattern Anal Mach Intell 33(10):1925–1937
4. Khorsheed JA, Yurtkan K (2016) Analysis of local binary patterns for face recognition under varying facial expressions. In: 2016 24th signal processing and communication application conference (SIU), May 2016, pp 2085–2088
5. Martinez AM (1998) The ar face database. CVC Technical Report. http://ci.nii.ac.jp/naid/10016836216/en/
6. Moorhouse A, Evans AN, Atkinson GA, Sun J, Smith ML (2009) The nose on your face may not be so plain: using the nose as a biometric. In: 3rd international conference on imaging for crime detection and prevention (ICDP 2009), pp 1–6
7. Mordini E (2014) Biometrics, pp 505–526. Springer, Netherlands, Dordrecht
8. Nambiar AM, Correia PL, Soares LD (2012) Frontal gait recognition combining 2d and 3d data. In: Proceedings of the on multimedia and security. MM&Sec '12. ACM, New York, NY, USA, pp 145–150. http://0-doi.acm.org.ujlink.uj.ac.za/10.1145/2361407.2361432

9. Oscs GC, Khoshgoftaar TM, Wald R (2014) Rotation invariant face recognition survey. In: Proceedings of the 2014 IEEE 15th international conference on information reuse and integration (IEEE IRI 2014), pp 835–840, Aug 2014
10. Thompson AF, Alese BK, Olofinlade FV (2013) Nose biometrics verification using linear object technique. In: 2013 Pan African international conference on information science, computing and telecommunications (PACT), July 2013, pp 182–187
11. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
12. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb
13. Zehngut N, Juefei-Xu F, Bardia R, Pal DK, Bhagavatula C, Savvides M (2015) Investigating the feasibility of image-based nose biometrics. In: 2015 IEEE international conference on image processing (ICIP), Sept 2015, pp 522–526
14. Zuo W, Wang K, Zhang D (2006) Robust recognition of noisy and partially occluded faces using iteratively reweighted fitting of eigenfaces. Springer, Berlin, Heidelberg, pp 844–851

Part IV
Middleware and Operating Systems

Implementation of a Monitoring System for the Measurement of Temperature, Flow Rate, and Fluid Pressure of Cooling Systems

Cheol-Hong Moon

Abstract In this study, a monitoring system was constructed so that changes in the temperature of the fluid that flows inside piping, pipe surface temperatures, flow rates, and pressure can be measured simultaneously and the measured data can be displayed on a screen to enable the observation and recording of changes in fluid conditions for the analysis of the characteristics of automobile cooling systems and various fluid systems. Two types of temperature sensors (contact type and noncontact type) were installed in the system to measure the cooling efficiency of the cooling system and the monitoring system was configured using a digital semiconductor pressure gauge and a turbine flowmeter. A small ATmega8 was used as a CPU for each embedded sensor system. In addition, 24-bit AD converters were used for accurate measurement and the serial communication mode was used to transmit data from individual sensor systems.

Keywords Measurement · Temperature · Flow rate · Fluid pressure
Cooling system

1 Introduction

Cooling systems are used to maintain the engine at appropriate temperatures for the enhancement of fuel efficiency and output. Most recent automobiles use a water-cooled system consisting of a cooling pump, radiator, and various connecting pipes. Following the recent trend toward automobile weight reduction, studies on integral modularization and weight reduction of such water-cooled cooling systems are in progress. In particular, studies on system weight reduction by changing the materials of water pumps and pipes of cooling systems to plastic, except for radiators, are in progress [1]. In addition, studies for the analysis of the characteristics of such cooling systems are in progress [2]. In this study, a monitoring system was constructed so

C.-H. Moon (✉)
Electrical & Electronic Engineering, Gwangju University, Gwangju 61743, South Korea
e-mail: chmoon@gwangju.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_20

that changes in the temperature of the fluid that flows inside piping, and plastic pipe surface temperatures, flow rates, and pressure can be measured simultaneously and the measured data can be displayed on a screen to enable the monitoring and recording of changes in fluid conditions for the analysis of the characteristics of the weight-reduced cooling systems and various fluid systems in their entirety.

To measure the cooling efficiency of cooling systems, two types of temperature sensors (contact type and noncontact type) were installed in the system so that the temperatures in various regions of the cooling system could be measured. Temperatures are physical quantities that can be measured relatively easily using various sensors such as resistance temperature detectors (RTD) [3], thermocouples, infrared measurement sensors, thermistors, and semiconductor temperature sensors. With regard to thermometers, studies on thermometers and studies for the efficient utilization of thermometers are in progress [4].

The flow meters used to measure flow rates in pipes include differential pressure flowmeters, area flowmeters, ultrasonic flowmeters, turbine flowmeters, and volumetric flowmeters [5]. In the case of a turbine flowmeter [6], a rotor is installed inside the flowmeter so that the rotor rotates when a fluid flows inside the pipeline in proportion to the velocity of the fluid and the rotational speed of the rotor is measured to obtain the flow rate of the fluid flowing through the pipe.

In this study, sensors suitable for the measurement of fluid characteristics inside automobile cooling systems were selected. As for temperature sensors, RTD pt100Ω sensors that can measure temperatures in a range of −40 to +120 °C, which is the range of automobile engine temperatures, and that show quick responses to temperatures were selected, and infrared thermometers were selected to measure plastic pipe surface temperatures. As for pressure gauges, electronic pressure gauges that can precisely measure low pressures (0–1.6 kgf/cm^2) were used. As for flowmeters, turbine flowmeters with a compact design that can precisely detect flow rates (10–160 LPM) and can endure changes in temperatures from low to high were used.

In this study, the system was implemented based on RS422 serial communication so that it could transmit measured data to a PC in real time. The thermometers, flowmeter, and fluid pressure gauge system were designed and fabricated so that the data from individual sensors could be measured in real time, and the measured data are transmitted and displayed on the PC screen in their entirety under the control of the designed GUI monitoring program.

2 H/W System Configuration

2.1 System Overview

Figure 1 shows the block diagram of the monitoring system. A contact type thermometer for the measurement of temperatures inside the cooling system and a noncontact type infrared thermometer for the measurement of cooling system surface

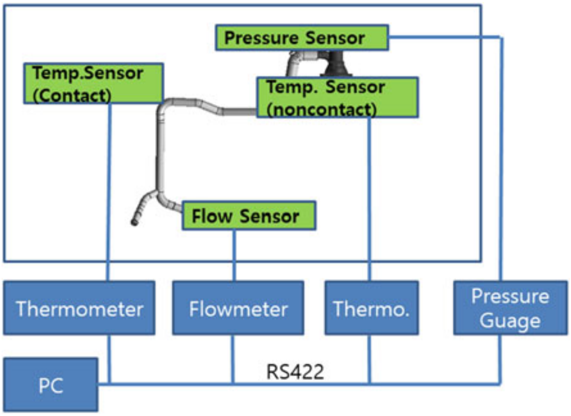


Fig. 1 Block diagram of the temperature, flow rate, and pressure monitoring system

temperatures were designed. A flowmeter for the measurement of flow rates of fluid passing through the piping was installed using a turbine flow rate sensor, and a pressure gauge using the semiconductor pressure sensor was designed so that fluid pressure inside the pipe could be measured. The thermometer indicated temperatures in units of 1 °C on an LCD, the flowmeter indicated flow rates of the fluid passing through the piping in units of L/min, and the pressure gauge indicated pressure in units of kgf/cm^2. In addition, the system was designed to transmit all data to the main control unit using serial communication so that changes in the temperature, flow rate, and fluid pressure of the cooling system could be continuously monitored, and it was also designed to display the transmitted data in a separate window on the monitor.

2.2 *Thermometer*

Figure 2 shows thermometer system circuit diagram. An embedded system for temperature measurement was made using an RISC microprocessor [7]. This system was configured with a 16 MHz clock unit, a system reset unit, a serial communication unit, an LCD display unit, a 24-bit AD conversion unit for an RTD temperature sensor interface, and an infrared temperature sensor unit.

The A/D conversion unit for contact type RTD temperature sensor has two 24-bit ADC channels. In Fig. 3, the serial output data stream of A/D conversion is 32 bits. The first 4 bits indicate completion of conversion, channel selection, conversion results, and output ranges. The following 24 bits are temperature data bits and indicate from MSB bit to LSB bit. The last 4 bits are not used.

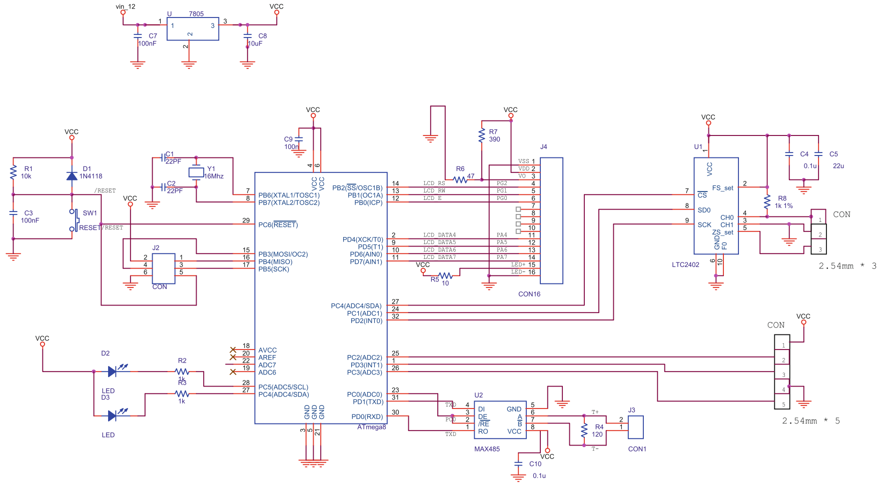


Fig. 2 Thermometer system circuit diagram

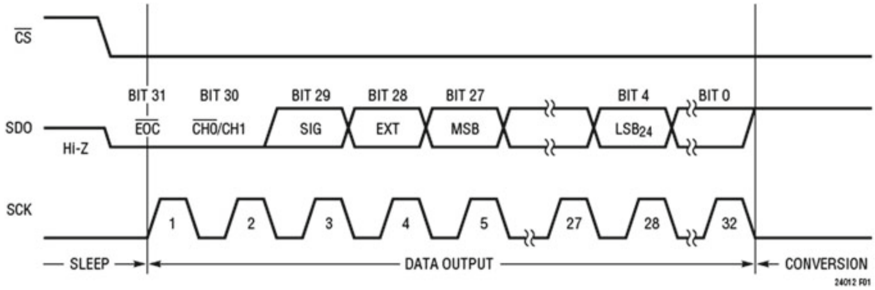


Fig. 3 Temperature sensor data timing

2.3 *Flowmeter*

The embedded system for flow rate measurement was also made using the same ATMega8 RISC microprocessor used in the thermometer, and was configured with a 16 MHz clock unit, a system reset unit, a power supply unit, a serial communication unit, an LCD display unit, an AD conversion unit, a pulse conversion unit, and an IO interrupt input unit. In this system, the pulses transmitted from the turbine flow rate sensor were amplified and transmitted to the IO of the processor and the amplified pulses were interrupted and counted to measure the flow rate.

2.4 Fluid Pressure Gauge

The system for measurement of fluid pressure was configured with the same clock and a circuit similar to those of the thermometer and the flowmeter for the synchronization of the entire system. However, in the AD conversion unit for pressure measurement, although the same AD converter as in the thermometer was used, a three-line system was used in the contact type thermometer, while a two-line system was used in the pressure gauge.

3 Control Program

3.1 RTD Temperature Conversion Program

For the measurement of temperature, the resistance value of the temperature sensor was converted into voltage in a range of 0–5 V and entered into the thermometer system. The accurate voltage value was calculated from input signals using the 24-bit AD converter and the resistance value of the temperature sensor was calculated. Since the AD conversion is in the form of the 24th power, the input voltage was obtained by dividing the resistance value by 4096 (2^{12}) two times and multiplying the results by 5 V. The resistance value of the RTD sensor was obtained by using the voltage divider equation for resistors. To mitigate the chattering of input signals, the output was obtained by averaging three inputs and the final temperature value was obtained from a simplified resistance table.

3.2 Flow Rate Measurement Program

The flow rates were measured using interrupts occurring at one-second intervals and external IO input interrupts. The pulses inputted from the outside were counted using the interrupt [EXT_INT1] void ext_int1_isr(void) and the one-second timer is operated using the interrupt [TIM1_OVF] void timer1_overflow(void). To calculate the flow rate of the fluid passing through the pipes per minute, the value of interrupts counted for one second was multiplied by 60 s and divided by the volume factor.

Serial data are transmitted from the PC in the order of device IDs, as shown in Table 1, and the relevant devices transmit their data from the hundreds digit to digits below the decimal point.

The initial screen for monitoring, which was designed so that temperatures can be displayed on the left top and bottom 2CHs. Temperatures ranging from −40 to 120 °C can be displayed, data for measurement 200 times can be displayed, and when data for measurement more than 200 times are entered, the oldest data are removed and the newly entered data are displayed as the last data. The top right field was

Table 1 Communication Protocol

(a) PC transmission		
Byte	Content	
0	'\r'	0D
1	'\n'	0A
2	Device ID	
	Temp. #1: 1	
	Temp. #2: 2	
	Pressure: 3	
	Flow: 4	
3	'\r'	0D
4	'\n'	0A
(b) PC receipt		
0	'\r'	0D
1	'\n'	0A
2	Device ID	
3	100's digit	ASCII
4	10's digit	ASCII
5	1's digit	ASCII
6	Below decimal point	ASCII
7	'\r'	0D
8	'\n'	0A

designed to display fluid pressure in a range of 0.0–1.6 kgf/cm^2, and the bottom right field was designed to display flow rates in a range of 0–160 L/M.

4 Experiment and Results

4.1 *Infrared Temperature Sensor*

Figure 4 Shows Infrared temperature sensor module. The infrared temperature measurement can measure temperatures in a range of −33 to 120 °C, and the precision of measurement is ±0.6 °C. The response speed of the sensor is 1 time/s. and the ratio of the distance of the object being measured to the distance of the focus is 1:1. The measured temperature values are displayed on the LCD in units of 1 °C. Based on the results of infrared temperature measurement, the measured temperature values drastically changed according to the separation distance between the object being measured and the infrared sensor. The ratio of the distance of the object being measured to the distance of the focus should be strictly maintained at 1:1, and because

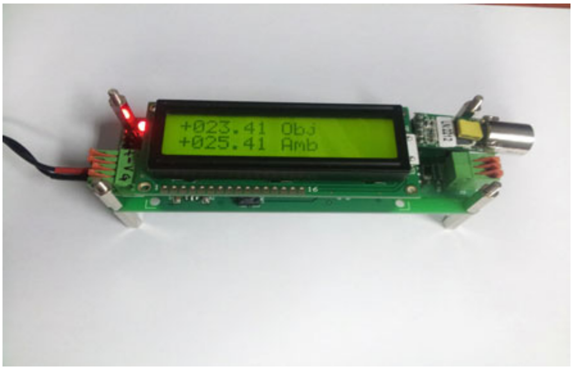


Fig. 4 Infrared temperature sensor

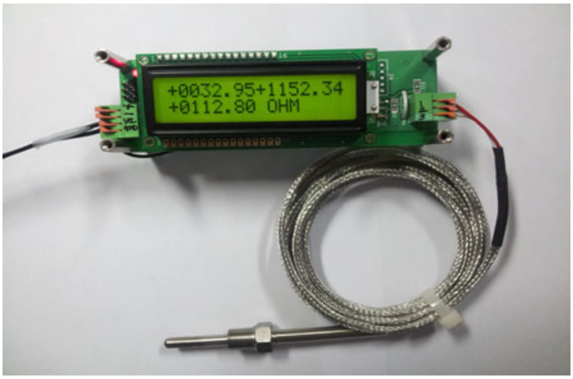
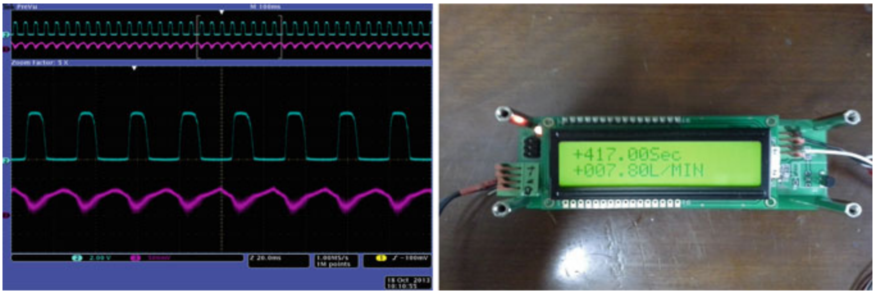


Fig. 5 RTD temperature sensor

of the characteristics of the sensor, objects at temperatures below −33 °C tended to show many errors. Figure 4 shows the infrared temperature sensor module.

4.2 RTD Temperature Sensor

Figure 5 shows the PT100Ω RTD sensor and the embedded system for temperature measurement. The RTD sensor distance is 1 M, and a three-line system was used. As a result of temperature measurement, 96.9 Ω at −7 °C and 130.6 Ω at +80 °C respectively. In the system where the PT100Ω sensor was used, the phenomenon of signal vibrations occurred. To minimize this phenomenon, replacement of the sensor with better response characteristics was reviewed, and an averaging algorithm that receives inputs many times and takes the average was adopted and used.



a) Waveforms inputted from the flow rate sensor

b) flowmeter

Fig. 6 **a** Waveforms inputted **b** flowmeter from the flow rate sensor

4.3 Flowmeter

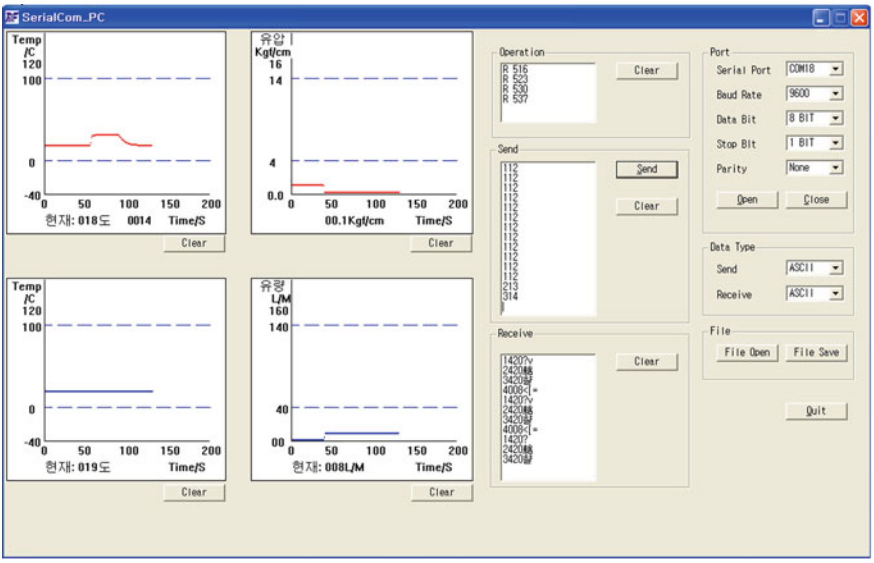
The half-wave rectifying signals inputted from the flow rate sensor are shown in Fig. 6a as pink signals, and are approximately 700 mV. These signals were inverted and amplified into digital waveforms so that they could be entered into the microprocessor. Figure 6b shows the fabricated flowmeter.

4.4 Fluid Temperature, Flow Rate, and Pressure Measurement

The data measured through the thermometers, flowmeter, and fluid pressure gauge were transmitted to the PC in the order set forth under the monitoring protocol. In the serial communication transmission, the IDs assigned to the thermometer #1, the thermometer #2, the fluid pressure gauge, and the flowmeter were 1, 2, 3, and 4, respectively. The communication baud rate was set to 9600BPS, the data bits were set to 8 bits, the stop bit was set to 1 bit, and the parity was set to none.

The contact type temperatures can be displayed in a range of −40 to 120 °C, and the noncontact type temperatures can be displayed in a range of −33 to 120 °C. The temperature value entered first is displayed on the very left side of the screen and the following values are displayed to the right.

Up to 200 measured data points can be entered, and when more than 200 data points have been entered, the data entered first are removed and all remaining data are moved to the left by 1 bit per data. In the fabricated thermometer, errors of up to 2 °C occurred before the averaging algorithm was applied, and errors of up to 0.4 Ω occurred due to errors in the resistance of the circuit. After applying the averaging algorithm and removing the resistance component of the line, the errors could be reduced to below 1 °C compared to a commercial industrial thermometer.



Operation

```
R 516
R 523
R 530
R 537
```

Send

```
112
112
112
112
112
112
112
112
112
112
112
112
213
314
```

Receive

```
1420?v
2420轄
3420잘
4008<[=
1420?v
2420轄
3420잘
4008<[=
1420?
2420轄
3420잘
```

Port

Serial Port: COM18

Baud Rate: 9600

Data Bit: 8 BIT

Stop Bit: 1 BIT

Parity: None

Data Type

Send: ASCII

Receive: ASCII

File

Fig. 7 Fluid monitoring screen

In addition, since the response speed of the RTD thermometer is low, the response speed should be increased in future experiments by using other types of sensors. The system was designed to be able to display fluid pressures in a range of 0–1.6 Kgf/cm^2. When the resultant values of the mechanical pressure gauge and the digital pressure gauge of the fluid measurement module were compared with each other, almost no error occurred at pressures below 1.6 Kgf/cm^2. The system was designed to be able to display flow rates in a range of 0–160 L/min. Figure 7 shows data for approximately 130 s. entered through serial communication. The flow rate is 8 L/min, the fluid pressure in this case is approximately 0.1Kgf/cm^2, and the temperatures of individual channels are shown to be 18 °C and 19 °C, respectively. Finally Fig. 8 shows the whole measuring system implemented.

5 Conclusion

In this study, a contact type thermometer, a noncontact type thermometer using infrared rays, a flowmeter using turbine flow rate sensors, and a fluid pressure gauge using semiconductor fluid pressure sensors were designed and fabricated. In addition, a GUI monitoring program that can transmit data from the individual fabricated measuring instruments to a PC using serial communication to display the data on a screen in their entirety and store the data in the PC was designed. The program was designed to display the various measured fluid data simultaneously on one screen so

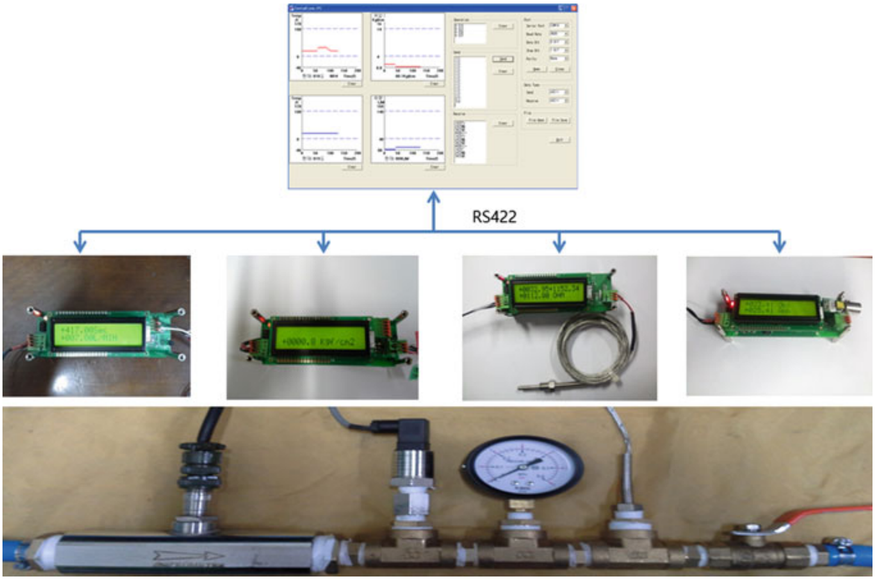


Fig. 8 Fluid measurement system implemented

that overall states of changes in the fluid could be seen at a glance and to shift the oldest data and enter the newest data at the end.

For this system, the range of temperature measurement was set from −40 to +120 °C to fit the experiments for a passenger car cooling system, and the temperature sensor was designed to be inserted directly into the fluid pipes to directly measure the fluid temperature. This system was designed to measure flow rates by receiving signals in the form of pulses using a turbine flowmeter and counting the signals and measuring up to 160 L/min. The fluid pressure gauge was designed to measure fluid pressures up to 1.6 Kgf/cm^2.

However, since the RTD type temperature sensor showed a low temperature response speed, the temperature sensor had to be replaced by a K type thermocouple with a high response speed, and since the output current of the semiconductor pressure sensor used in the pressure gauge was very low, an IC with low input current had to be selected in the process of interfacing. There was no particular problem with the flow rate measurement using the flowmeter because flow rates were measured using two interrupts: one timer interrupt and one pulse counter interrupt. However, a lot of time was spent in the process of communication to transmit measured data. Therefore, when designing the system later, a high speed CPU, which is faster, should be used.

Acknowledgements This Study was conducted by research funds from Gwangju University in 2018.

References

1. Hanil Tube Co (2015) The development of plastic cooling system applied water injection technology (WIT) for power train. The technical Report of Ministry of Trade, Industry and Energy
2. Sin Y (2013) Characteristics of pressure and flow rate of the cooling water in accordance with the water pump speed in a passenger car diesel engine. In 2013 annual conference and exhibition of the Korean society of automotive engineering, pp 277–278
3. Kim Y-G, Kim SH, Yang I (2009) Measurement of the time constant of industrial platinum resistance thermometers. J Korean Soc Precis Eng 26(11):41–46
4. Kim K-B, Lee H-K (2002) A study on the dynamic characteristics of thermowell by flow fluid using semi-empirical method. In: The conference of the Korean society for noise and vibration engineering, pp 147–152
5. DoHyung LEE (2008) Research and development of flow meter. KSFM J Fluid Mach (Korean Society for Fluid Machinery) 11(1):95–99
6. Kim J, Ko S (2003) Numerical study of three-dimensional flow through a turbine flow meter. J Korean Soc Fluid Mach 6(1):44–50
7. Atmel (2013) 8-bit Atmel with 8KBytes in system programmable flash ATmega8. Technical reference manual

Part V
Security and Privacy

A Review of Threat Profiling Techniques for Use in Concealed Weapon Detection Systems

Kudzaishe Mhou and Dustin van der Haar

Abstract In this paper, we discuss different types of threat assessment approaches for use in surveillance systems. We look at each method separately from others with the aim of showing the evolvement of the technology. A detailed description and an in-depth look at current research are conducted in order to give a more concise review. Doing an in-depth individual analysis of each method allows for a better comparison approach with other methods. We also look at the advantages and disadvantages of each approach and compare it with other methods in order to get a better sense of the performance of the technique as compared to others. A discussion on a benchmark comparison of these methods is also given and a few recommendations are presented along with our proposed approach.

Keywords Concealed weapon detection (CWD) · Close circuit television (CCTV)

1 Introduction

In recent years, the detection of individuals concealing weapons has become a task of vital importance with the increase in acts of terror. Terrorism attacks have been on the rise with various types of weapons having been used to cause devastation within societies, from school shootings by teenagers to crowd stabbing by elderly people [1]. These incidents can be used to show the complexity around developing an ideal weapon detection system since there is no single criteria that is common in these incidents. As such research efforts have over the years focused on detecting anomalies in either magnetic field distortions or techniques that are can see through clothing [2].

K. Mhou · D. van der Haar (✉)
Academy of Computer Science and Software Engineering, University of Johannesburg, Cnr University Road and Kingsway Avenue, APK Campus, Johannesburg 2006, South Africa
e-mail: dvanderhaar@uj.ac.za

K. Mhou
e-mail: 216074404@student.uj.ac.za

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_21

The aim of these techniques is to enable law enforcement to catch criminals before they commit crimes. However, with the ever-changing technology, civil rights and types of weapons it has become a mammoth task to detect individuals concealing weapons. Most acceptable techniques used in public spaces such as metal detectors fail when detecting more modern weapons such as 3D printed guns or ceramic weapons [3].

The use of the Internet has also resulted in the increase in violent crimes since criminals from all over the world are able to share ideas on how to evade the law and also sell weapons without ever crossing state lines. Very little research has been done on the effects of concealing a weapon. Effects such as the change in gait and change in behavioral traits. In this paper, we look at computer vision based methods and their progress in the past decade. We compare these methods using a benchmark and discuss each method's advantages and disadvantages along with how it can be improved, we also highlight trends and promising techniques in concealed weapon detection.

This paper is structured as follows, at the beginning of Sect. 1 we give a brief introduction and explain the problem at hand, Sect. 2 we discuss different types of vision-based methods for threat assessment followed by a comparison of methods using a benchmark in Sect. 2. Section 3 discusses our proposed approach for reviewing techniques discussed in this review. Lastly, Sect. 4 outlines a discussion of concealed weapon detection and gives a conclusion to the discussion at hand whilst giving recommendations on ways to improve these techniques.

1.1 Problem Statement

A great need exists for a concealed weapon detection system that can be used to detect not only concealed weapons but also individuals concealing the weapons in a security surveillance system. Current camera-based security surveillance systems are not capable of achieving these goals. Integrating common techniques such as X-ray imaging or millimeter wave scanners in these systems presents problems of privacy concerns and scalability. Some of these systems that make use of X-ray or millimeter wave tend to reveal hidden anatomical features which result in privacy concerns [4]. Therefore, there is a need for a novel approach that is cost-effective, has no privacy concerns, easily scalable and can be integrated into existing systems.

2 Concealed Weapon Detection

In this section, we look at some vision-based systems that are used for concealed weapon detection. We look at research that has been conducted over the past decade in this field. We also give a discussion around the performance of these systems and their ability in assisting law enforcement to prevent crimes.

2.1 Infrared Based CWD

The use of infrared in CWD is a technique that most researchers have raised skepticism about over the years. The skepticism is mostly due to the power of waves emitted by infrared sensors. Such waves are not powerful enough to penetrate most material like clothing [5]. Such an inability presents a challenge since detection of concealed weapons always happens under a piece of material. The theory behind the use of infrared is that the human body and the weapon being concealed have a temperature difference which results in the weapon glowing darker when viewed under an infrared sensor. However, infrared based weapon detection systems fail to scale in real-world scenarios because of factors such as differences in environments and distance of detection.

Over the years research conducted in using infrared for CWD has shown that in some instances infrared can detect weapons concealed under clothing. Most sensors used for infrared red imaging have a resolution of 1° which allows them to differentiate different objects. Research conducted by Mahadevi and Shridevi [5] highlighted that infrared could be used for concealed weapon detection. However, the results from the research showed that the method could only work for individuals with loose or tight clothing, Xue et al. [6] also came to a similar conclusion in their research. The results also showed that certain clothing areas that were are in contact with the human body have a similar color as the parts of clothing where a weapon is concealed making it difficult to accurately detect a weapon [6]. To solve these problems some researchers have suggested the fusion of techniques when using infrared [5, 6]. However, research conducted shows that the results of some fusion techniques are similar to those of using infrared alone. However, adding a human operator to work together with such systems improves detection rates as it is easier for human operators to spot the weapon [7].

2.2 Weapon Detection Based on Image Fusion

The combination of techniques is an approach that aims at leveraging the advantages of different techniques in order to improve accuracy rates when using a single fused system. Over the years a number of image fusion techniques have been implemented in weapon detection, fusion techniques such as Xray and metal detectors among others.

Researchers Zhang and Blum implemented a region-based system for weapon detection [8]. In their research, it was noted that using multiple images from different sensors significantly reduces error rates. The same conclusion was also highlighted in a research conducted by Xue and Blum where image fusion techniques were used to identify a weapon concealed underneath clothing, the difference being that in this research a single sensor was used [9]. The researchers used a wavelet-transform-based approach in their fusion algorithm where they combined aspects of feature-level and pixel-level fusion [8]. The algorithm looked at regions of an image as compared to singular pixels, it allowed the fusion algorithm to become less affected by noise and blurring effects that are often introduced when dealing with pixel-based image fusion algorithms.

The proposed algorithm was tested on multiple pairs of images and the results illustrated that the proposed algorithms worked well for both images from the same sensor and those from different sensors. The algorithm was able to identify concealed weapons underneath clothing [8]. In the paper by Xue and Blum a color image fusion algorithm was proposed which used a normal RGB image and an IR image to detect a weapon concealed underneath clothing. The research showed that the proposed algorithm was able to reveal the concealed weapon hidden underneath clothing and also the researchers illustrated that the algorithm could potentially be used to reveal weapons hidden in other materials such as bags [9].

The fusion of techniques has proven to be one of the most promising effective ways of detecting concealed weapons however most fusion techniques have been around the fusion of algorithms. Very few researchers have attempted to fuse existing CWD techniques with biometrics techniques. The lack of research in using biometric techniques in CWD has resulted in a gap in accuracy rates between biometric techniques and most CWD algorithms.

2.3 Threat Profiling Using Gait Analysis

The idea that the way one walks could be used as a threat assessment model is an idea that causes much debate in the scientific community. Nair et al. illustrates showed how artificial neural networks were used in developing a model for threat assessment based on gait. The research showed the possibility of using deep neural networks in threat profiling, by analyzing gait patterns as shown by the 60.32% accuracy attained during classification [10].

Deep belief neural networks were used in the research to extract motion and image features, the use of these neural networks assisted in the investigation of the inverse kinematic model [10]. The deep belief neural networks helped reduce the dependency on tracking mechanism during classification stages. When the same algorithm used in the research was tested without the deep belief neural networks only an accuracy of 52% was reached. The researchers also obtained a 59.36% accuracy when using the joint angle trajectories computed from SURF-based IRKF method. This research showed that analyzing gait for threat profiling has potential and if improved more accurate results could be obtained [10].

The use of gait in threat profiling is a promising technique, with advantages such as difficulty to camouflage and non-invasiveness. Stevenage et al. [11] showed that gait can be used as a recognition measure for individuals by analyzing gait signatures from videos. Accurately identifying gait allows for an opportunity to classify gait based on different criteria. Further research conducted by Jain et al. [12], highlighted that gait can be classified into cycles i.e. the time gap between consecutive instances of initial foot-to-floor contact for the same foot also known as heel strike. Using these research findings, a number of researchers have begun investigating the use of gait in threat profiling or assessment.

Razali and Manaf [13] developed a model to use when identifying threats using motion capture. In the research Principle Component Analysis (PCA) and Euclidean distance were used. The data captured through motion capture was processed using PCA to obtain feature vectors. These feature vectors were used as input for the Euclidean distance algorithm, which was used to determine similarities between datasets. Several findings were presented in the research such as the most significant movement pattern to distinguish gaits. The pattern found to be most effective is the pattern produced by ankle movement.

Futhermore BenAbdelkader and Davis [14] showed that using gait and other body movements can be effective in identifying people carrying objects. This discovery has many applications in threat profiling such as detecting suicide bombers and individuals concealing weapons. During the research one single stationary camera was used from a distance to detect the individuals carrying objects. The detection of these individuals was achieved via correspondence-free analysis of binary shape features that take into account the pendula like the motion of both an individual's legs and arms. A good detection rate of 85% was reached and a false alarm Arte of 12%.

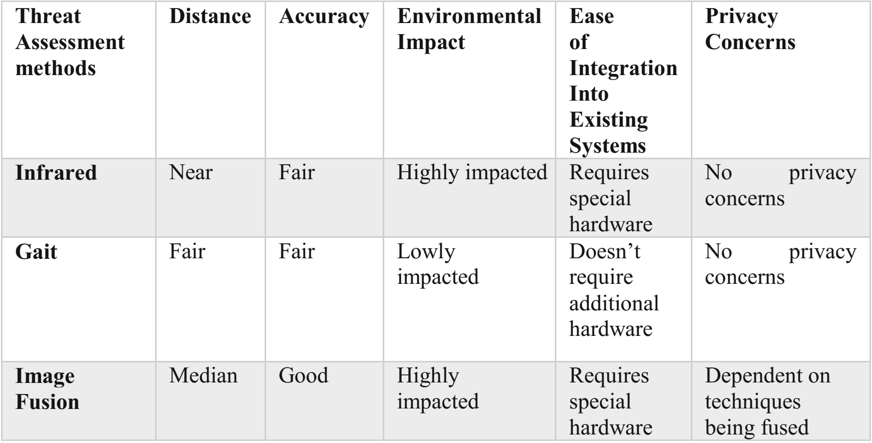
3 Methodology for Reviewing Techniques

In order to have a review that is less biased, we look at published research that supports the techniques we are reviewing. We look at how these techniques were implemented and the results which were achieved. Parameters such as distance for detection, environmental impact, accuracy, ease of integration in existing systems and privacy concerns are also considered in reviewing each technique. An analysis is given based on average benchmarks of these systems. By using average benchmark results we hope to eliminate bias towards one research.

4 Comparison/Benchmark

In this paper, we have looked at a small number of methods used in weapon detection. Our aim is to expose current methods, promising methods, highlight trends and give recommendations to the scientific community based on the literature review. To begin this section, we go through each benchmarking parameter and give an explanation of why such a parameter is justified to be a part of the benchmark metric.

Metrics such as the environmental impact on a system are vital for assessment because of different system performance in different environments. The assessment gives an insight of how the technique performs in real-world environments.



Threat Assessment methods	Distance	Accuracy	Environmental Impact	Ease of Integration Into Existing Systems	Privacy Concerns
Infrared	Near	Fair	Highly impacted	Requires special hardware	No privacy concerns
Gait	Fair	Fair	Lowly impacted	Doesn't require additional hardware	No privacy concerns
Image Fusion	Median	Good	Highly impacted	Requires special hardware	Dependent on techniques being fused

Fig. 1 The figure shows the average scores for each metric per technique

Information about how accurate the technique is can provide an insight into areas of improvement and show how the technique performs. However, accuracy alone does not give a complete insight of the performance of a technique because accuracy rates depend on environments. It is important to note that these metrics collectively contribute to the adoption of a system into the real world.

It is also vital that each technique is assessed whether it violates privacy laws. Looking at such an assessment allows for a more in-depth analysis of a technique. Also looking at how such techniques can integrate into current systems can give us a brief overview of how the technique could be easily adopted into existing systems. Making use of these five metrics we assess each technique and give a justification of each score assigned to a technique. Figure 1 illustrates our results.

4.1 Discussion

In this research, we gave a descriptive average as compared to a numeric calculation because in some of the results in the research reviewed in this paper were not presented as a numeric value rather they were presented in a descriptive manner. By giving a descriptive average we aim to quantify the score within a range rather than settle on a specific numeric value.

Looking at the distance of detection in Fig. 1 we assigned scores into three namely near, median and far. Near meaning a distance of fewer than 10 m, median less than 15 m but above 10 m and far any distance above 15 m. The distance of detection is one of the most important properties of a CWD system as it can provide time for law enforcement to react to criminals. Research conducted by BenAbdelkader and Davis show that the gait based techniques are capable of performing detection at a considerable distance. Techniques such as infrared and image fusion techniques are rather limited when it comes to the distance of detection due to the sensors used in capturing samples and the wavelength.

In terms of environmental impacts, it is noticeable that gait based techniques have a low chance of being impacted by environmental changes. We attribute this to the fact that gait systems do not rely on special sensors other than normal cameras to perform detection, unlike infrared or some image fusion techniques. Such special types of equipment such as infrared sensors are sensitive to certain environments which impact their detection ability.

In Fig. 1 image fusion techniques showed the highest accuracy. Fused techniques yield good accurate results because of the system's ability to leverage multiple technique's advantages. However, even though fusion techniques have a high accuracy privacy concerns are still an open problem in these systems since they are capable of seeing through clothing and reveal hidden anatomical feature. Techniques such as gait are less prone to these challenges which make them much more desirable for in threat profiling systems in the public space.

Current systems in place for threat assessment such as CCTV do not have a way of threat assessment without human intervention. The number of CCTV cameras around the world make it a mammoth task to replace these CCTV systems with ones that are capable of threat assessment. However, developing methods that can integrate seamlessly into existing systems is an objective which threat assessment systems ought to have. Techniques such as gait based system discussed in this paper have a greater chance of a seamless integration due to their ability to work with different types of cameras despite their resolution quality. Despite challenges such as low accuracy rates gait based threat profiling methods seem to be the most favorable techniques for threat profiling and weapon detection. What the gait techniques lack in accuracy they compensate in ease of integration, distance and less environmental impacts on the system. Thus, we suggest that gait is the most favorable technique that has promising findings and we feel that if more research is focused on using gait for weapon detection the accuracy could greatly improve.

4.2 Recommendations

From the review of these techniques, we can see a growing trend within the scientific community to use vision-based methods in the detection of concealed weapons. However, problems such as adaptation to different environments still remains an open problem. Most techniques discussed in this research have attempted to solve this problem, however, there still remains a gap. As a recommendation, we suggest fusion of gait with techniques such as infrared to improve the detection rates.

4.3 Conclusion

In this paper, we set out to look at the recent methods of threat assessment. We specifically focused our literature review on studying some of the most promising ways of threat assessment. We looked at the use of image fusion techniques, gait recognition, and infrared based techniques. Over the years the research shows that there has been a significant interest by many researchers in the field of threat profiling. The increase in research interest has resulted in some methods becoming more accurate than others because of improvements in such methods [2]. The fusion of methods has also shown great potential in threat assessment and weapon detection.

Looking at all the research discussed in this paper, we can conclude that these techniques have a good chance of solving the issue of concealed weapon detection. However, we would like to propose a fusion of gait and infrared techniques in creating a threat assessment model since each one of them has an advantage that can be used to boost the ability of the other.

References

1. Bonanno CM, Levenson RL Jr (2014) School shooters: history, current theoretical and empirical findings, and strategies for prevention. Sage Open 4(1):2158244014525425
2. Agurto A, Li Y, Tian GY, Bowring N, Lockwood S (2007) A review of concealed weapon detection and research in perspective. In: 2007 IEEE international conference on networking, sensing and control. IEEE, pp 443–448
3. Parande M, Soma S (2013) Concealed weapon detection in a human body by infrared imaging. Int J Sci Res 4(9):182–188
4. Roomi MM, Rajashankari R (2012) Detection of concealed weapons in X-ray images using Fzz K-NN. Int J Comput Sci Eng Inf Technol (IJCSEIT) 2(2):187–19
5. Parande M, Soma S (2015) Concealed weapon detection in a human body by infrared imaging
6. Xue Z, Blum RS, Li Y (2002) Fusion of visual and IR images for concealed weapon detection. In: Proceedings of the fifth international conference on information fusion, 2002, vol 2. IEEE, pp 1198–1205
7. Cho S, Tin N (2010) Using infrared imaging technology for concealed weapons detection and visualization, Fukuoka, 2010
8. Zhang Z, Blum RS (1997) Region-based image fusion scheme for concealed weapon detection. In: Proceedings of the 31st annual conference on information sciences and systems, pp 168–173
9. Xue Z, Blum RS (2003) Concealed weapon detection using color image fusion. In: Proceedings of the 6th international conference on information fusion, vol 1, pp 622–627
10. Nair BM, Kendricks KD (2016) Deep network for analyzing gait patterns in low resolution video towards threat identification. Electron Imaging 2016(11):1–8
11. Stevenage V, Nixon MS, Carter JN, Cunado D, Huang PS (1999) In: Jain A, Bolle R, Pankanti S (eds) Biometrics: personal identification in a networked society. Kluwer Academic Publishing, Dordrecht, pp 231–250

12. Jain AK, Bolle R, Pankanti S (eds) (1999) Biometrics: personal identification in a networked society. Kluwer Academic Publishing, Dordrecht, pp 231–250
13. Razali NS, Manaf AA (2011) Gait analysis for criminal identification based on motion capture
14. BenAbdelkader C, Davis L (2002) Detection of people carrying objects: a motion-based recognition approach. In: Proceedings fifth IEEE international conference on automatic face and gesture recognition, 2002. IEEE, pp 378–383

Performance Comparison of Some Addition Chain Methods Based on Integer Family

M. F. A. Kadir, M. A. Mohamed, R. Mohamad, M. Mamat and A. Muhammed

Abstract A generalized version of an addition chain problem, in which one must find a chain that simultaneously satisfies a sequence on integer in ascending order, is NP-complete. There is no known algorithm which can calculate an optimal addition chain for a given number with any guarantees of reasonable timing or small memory usage. Several methods were introduced to calculate relatively short chain and they are most used to support scalar multiplication operation tailored to limited computational resources in elliptic curve cryptography. In reality, one method is no better than the other except on certain occasions and only for specific integers. In this studies, we evaluate some existing addition chain methods against each other for their competitive performance by categorizing integers into various groups as the input. This result can be used as a benchmark for which method is suitable in which condition anticipated.

Keywords Addition chain · Elliptic curve cryptography · NP-complete
Heuristic method · Composition method

1 Introduction

Nowadays, information security is at the greatest importance in which communication over open networks and storage of data in digital form plays a key role in daily life. The science of cryptography provides efficient tools to secure information.

M. F. A. Kadir (✉) · M. A. Mohamed · M. Mamat
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia
e-mail: fadzil@unisza.edu.my

R. Mohamad
Department of System and Networking, Universiti Tenaga Nasional, Kajang, Malaysia

A. Muhammed
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Seri Kembangan, Malaysia

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_22

Cryptography is defined as the study of mathematical techniques related to aspects of information security such as confidentiality, data integrity, entity identification, and data origin authentication. Due to their tampered resistance, cryptosystems are often implemented on constraint memory devices such as smart cards. For such cases elliptic curve cryptosystems (ECC) [1, 2] is considered to be the most appropriate. The ECC exploited the discrete logarithm problem on a general elliptic curve that has no subexponential time solution. The major advantage of ECC is that a small key of size 160-bits can provide comparable security level with other cryptographic standards such as RSA of 1024-bits but with much faster and and more efficient execution.

The basic operation of ECC involves scalar and multi-scalar multiplication. Addition chain method has been widely used to improved efficiency of large number operation such that found in ECC. An addition chain (AC) for a number n is a sequence, such that each new member is the sum of two earlier (not necessarily distinct) ones. The length of an AC for an integer n is calculated as the number of terms other than the first one. The number of operations is directly proportional to the number of terms. This way, efficiency can be achieved if we can have shorter chain that is, the number of operations can be reduced by shortening the sequence.

Many AC method have been introduced. This is due to finding the optimal chain for a set of numbers was proven to be NP-complete [3]. These methods aimed at generating chain closest to optimal value. AC methods can be grouped into two major family, heuristics and metaheuristics. One method is known to be no better than the other except on certain occasions for certain groups of integers. However, there is lack of performance evaluation in term of length generated to compare the efficiency of different methods. Therefore, some performance analysis focusing on finding the shorter length need to be done. In this study, we conduct performance evaluation by comparing the length of selected AC for heuristic methods to find the best methods (with shortest length) by focusing on comparing seven methods, evaluating integer from 2 to 180,000.

The remainder of this paper is structured as follows; Sect. 2 summaries the review of literature. Section 3 states the materials and methods used in this project. These will be followed by the results and discussion in Sect. 4 and conclusion in Sect. 5.

2 Related Works

There are many AC methods that have been introduced in order to find sub-optimal solution. The literatures divide them into two categories that are meta-heuristics and heuristics [4]. A meta-heuristics is an iterative master process that guides and modifies the operations of subordinates heuristics to efficiently produce high-quality solutions. Meta-heuristics support decision making with robust tools that provide high quality solution to important application in business, engineering, economics and science. Common target of meta-heuristics are to solve optimization problem usually known by their complexities and it is inspired by analogies related to other factor such as natural, chemical, biological, electrical and thermal.

In solving AC problem, many meta-heuristics methods have been proposed. Genetic algorithms was first used by [5], allowing one-point crossover and uniform mutation, followed by [6] with the use of two-point crossover together with a local search mutation operator and a repair mechanism built within the initialization of the population. The most notable one, which employ a representation based on factorial number system together with neighborhood functions and distribution functions is credited to [7]. Various other methods such as ant colony optimization [8], artificial immune system [9], population-based optimization [10], and simulated annealing [11]. However, the most notable one is that of an evolutionary programming (EP) [12]. EP simulates evolution at species level. Therefore, no crossover operator is employed. In this proposed approach, an individual is represented at genotype level, that is an individual is a feasible AC. The fitness value of each individual is the length of the AC. Therefore, shorter strings are preferred. An advantage of the EP algorithm comprises the solution encoding with suitable fitness function and initial population, a mutation operator, and the survivor selection mechanism. These elements are easy to implement comparing to operators such as crossover and parent selection found in genetic algorithm. However, there are a few issues with meta-heuristics techniques such as consume a lot time to get optimal results, dependencies of specific cases for good result and complexities in implementation.

Heuristic is a much simpler techniques which seeks good solutions at a reasonable computational cost. Heuristic is designed and tuned for some specific problem. Heuristics are criteria, methods, or principles for deciding which among several alternative courses of action promises to be the most effective in order to achieve some goal. In broad, we can categorized these methods into a few different ways. By looking into the way of input representation we have binary or m-ary methods by using different radix, and unsigned and signed methods by allowing both positive and negative integer values. Either way, the idea is to reduce the number of operations that are addition and doubling, as much as possible. This can be done via manipulating the representation such that the digit representing the operation is reduced to the digit unpresented such that found in Binary Method (BM) [13]. Moreover, the location and adjacency of the digit can also be impactful for some methods such as Non-Adjacent Form (NAF) [14], and Complementary Recoding (CR) [15], Decomposition Method (DM) [16], Composition Method (CM) [17], Signed Decomposition Method (SDM) [18], and Signed Composition Method (SCM) [19].

3 Materials and Method

In general the studies of AC can be represented by a framework in Fig. 1. This section discusses how do we conduct the experiment and the tools required. The performance metric used is the length of the generated AC. The chosen methods are Binary Method, Non Adjacent Form, Complementary Recording, Decomposition

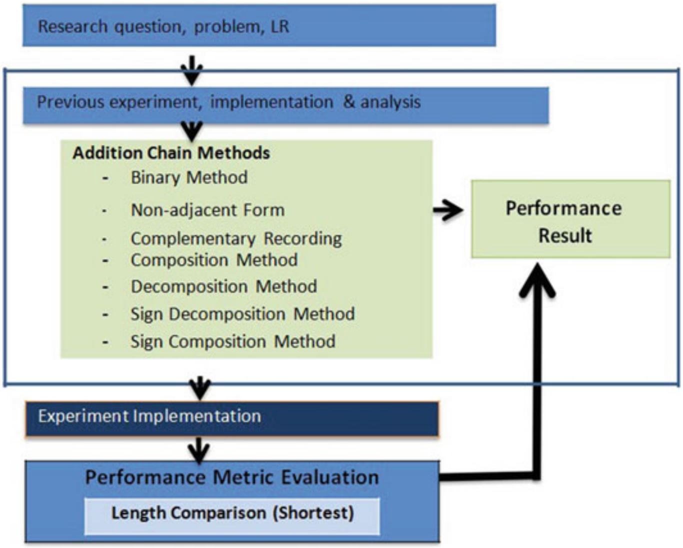

Research question, problem, LR
Previous experiment, implementation & analysis
Addition Chain Methods
- Binary Method
- Non-adjacent Form
- Complementary Recording
- Composition Method
- Decomposition Method
- Sign Decomposition Method
- Sign Composition Method
Performance Result
Experiment Implementation
Performance Metric Evaluation
Length Comparison (Shortest)

Fig. 1 Investigative framework

Method, Composition Method, Signed Decomposition Method, and Signed Composition Method.

We implemented these algorithms based on that of suggested by the original articles and perform the performance measurement under Dev C++ environment. The GNU Multiple Precision Arithmetic Library (GMP) is a free library for arbitrary-precision arithmetic, operating on signed integers, rational numbers, and floating point numbers is used to support large integer operation. GMP has a rich set of functions, and the functions have a regular interface. Besides that main target applications of GMP are cryptography applications and research same as in our project. To install the GMP library a few tool should also be installed which are MinGW 3.4.2—used as interfaced, GCC and GCC++, and MySYS 1.0.10. Meanwhile, for graphing and analysis the GNUPlot 4.6 and Microsoft Excel 2010 are used respectively.

This whole experiment can be divided into 3 phases. First, we develop all the seven algorithms and measure their performance for integers from 1 to 180,000. An overall from this experiment, we can see the distribution of length for each integer. Second, we do the comparison using four group of test which are:

Test 1: We compare every two methods, totalling 21 combinations altogether, to find out how many wins (shorter), looses (longer) and draws (equal) in terms of the length of AC measured for every integer in each block.
Test 2: By splitting integers n into blocks such that $2^N + 1 \leq n \leq 2^{N+1}$ specified by N from 3 to 17, we compute an average length for each block.
Test 3: By separating odd $(2k + 1)$ from even $(2k)$ integers, we study the distribution of an AC generated by each method.

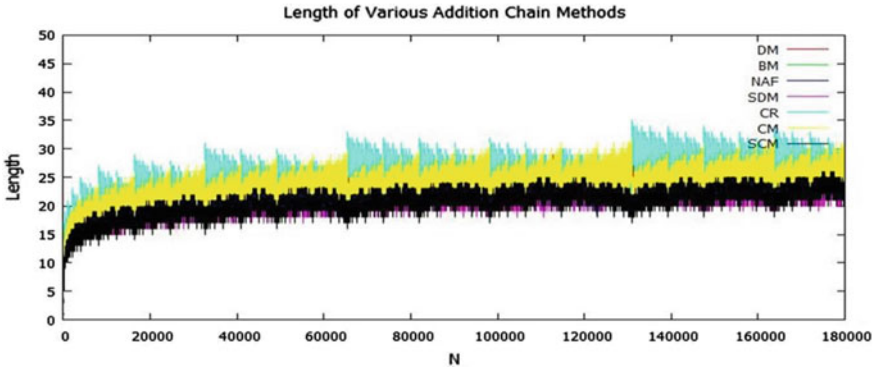


Fig. 2 Distribution length of 7 methods

Finally, we plot the graph for every test done to observe the distribution and comparison then analyze the result.

4 Results and Discussion

In this section, we discuss the results that we have been obtained from our three experiment setup.

For Test 1, Fig. 2 shows the distribution of the length of AC produced by the seven different methods. Observably, the lengths for all methods are increasing when the integer are growing but the rates of growth are fairly small. Table 1 shows the list of comparisons between methods for 21 combinations. These comparisons calculate the wins, loose and draw in term of the length of AC. The method with more wins (shorter chain) is considered as winner. The more the method wins the shorter the method generated the length. The SCM has wins all the comparison against the other methods whereas, CM and BM produces chain of having the same lengths.

For Test 2, the average length of AC for each block specified by the value of N are calculated by adding all the lengths for every integer and divide by N. Table 2 shows the result for $3 \leq N \leq 17$. We observed that for $N \leq 8$, SDM produces the shortest average among all methods, whereas for $N > 8$, SCM seems to outclass all other methods.

From the Test 1 and Test 2, the methods can be ranked as SCM (shortest), SDM, NAF, DM, BM and CM, and CR. In both tests, we found out SDM and SCM methods are competing between each other in term of performance. The length of SCM becomes shortest when the integer number growth.

In Test 3, by separating even from odd integers, we plot graphs representing the length of AC produced by each method. For odd integer, Fig. 3 shows that SDM and SCM are competing to each other to become the method with the shortest AC. For

Table 1 Length comparison

Methods	Wins	Loose	Draw	Results
DM versus BM	77647	41789	60564	DM
DM versus NAF	28829	113899	37272	NAF
DM versus CR	129105	31093	19802	DM
DM versus SDM	0	112992	67008	SDM
DM versus CM	77647	41789	60564	DM
DM versus SCM	18194	125631	36175	SCM
BM versus NAF	13010	122471	44519	NAF
BM versus CR	108370	53573	18057	BM
BM versus SDM	12786	140287	26927	SDM
BM versus CM	0	0	180000	-
BM versus SCM	0	133111	46889	SCM
NAF versus CR	150918	0	29082	NAF
NAF versus SDM	41942	72350	65708	SDM
NAF versus CM	122471	13010	44519	NAF
NAF versus SCM	0	48929	131071	SCM
CR versus SDM	5317	163572	11111	SDM
CR versus CM	53573	108370	18057	CM
CR versus SCM	0	164937	15063	SCM
SDM versus CM	140287	12786	26927	SDM
SDM versus SCM	49898	50296	79806	SCM
CM versus SCM	0	133111	46889	SCM

Table 2 Average lengths

$2^N + 1 \leq n \leq 2^{N+1}$	DM	BM	NAF	SDM	CR	CM	SCM
$N = 3$	4.5	4.625	4.875	4.5	5.75	4.625	4.5
$N = 4$	5.9375	6.0625	6.125	5.75	7.125	6.0625	5.8125
$N = 5$	7.34375	7.53125	7.46875	7.125	8.5625	7.53125	7.125
$N = 6$	8.71875	9.01563	8.78125	8.42188	10.0313	9.01563	8.45313
$N = 7$	10.1797	10.5078	10.1172	9.76563	11.5156	10.5078	9.78125
$N = 8$	11.6328	12.0039	11.4453	11.1016	13.0078	12.0039	11.1133
$N = 9$	13.0898	13.502	12.7793	12.4473	14.5039	13.502	12.4453
$N = 10$	14.54	15.001	14.1113	13.7793	16.002	15.001	13.7783
$N = 11$	16.0073	16.5005	15.4448	15.1138	17.501	16.5005	15.1113
$N = 12$	17.4746	18.0002	16.7778	16.4514	19.0005	18.0002	16.4446
$N = 13$	18.9456	19.5038	18.1149	17.7887	20.5072	19.5038	17.7815
$N = 14$	20.4062	20.9957	19.4403	19.1133	21.994	20.9957	19.107
$N = 15$	21.8826	22.5	20.7778	20.4524	23.5001	22.5	20.4445
$N = 16$	23.3583	24	22.1111	21.7878	25	24	21.7778
$N = 17$	24.8326	25.5	23.4445	23.4445	26.5	25.5	23.1111

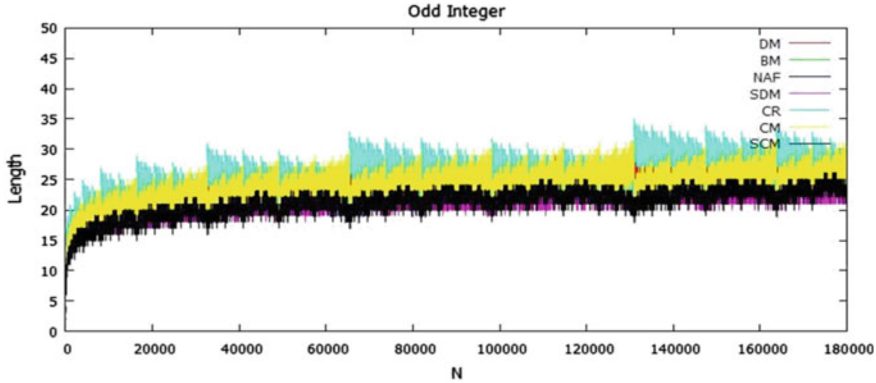


Fig. 3 Distribution of length for odd integer

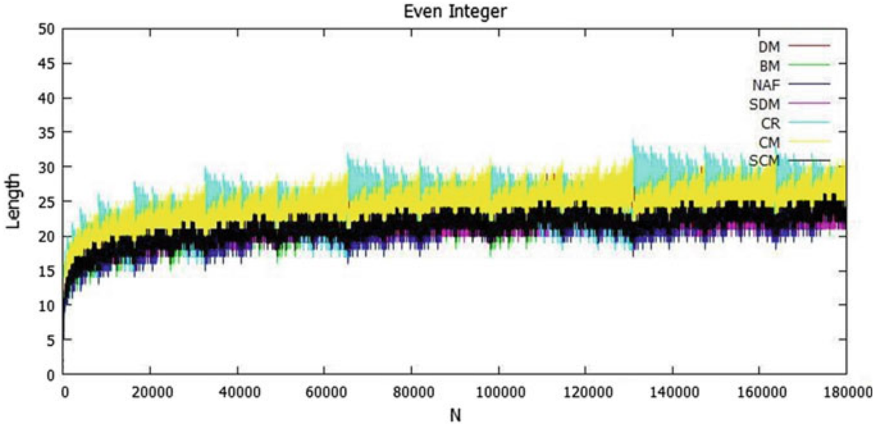


Fig. 4 Distribution of length for even integer

even integers, Fig. 4 shows that NAF method has contributed to the production of the shortest AC alongside with SDM and SCM.

5 Conclusion

Competitively, SDM and SCM generate the shortest length. Both methods use rules as their basis. However, SCM has outclassed SDM for large integers and therefore is more recommended for cryptographic implementation which utilizes large integers.

Acknowledgements This study is funded by Ministry of Higher Education Malaysia (FGRS/1/2017/ICT03/UNISZA/02/1(RR228)). Special thanks to University of Sultan Zainal Abidin for providing facilities to develop and evaluate the proposed framework.

References

1. Mohamed MA (2014) A survey on elliptic curve cryptography. Appl Math Sci 8(153–156):7665–7691
2. Koblitz N (1987) Elliptic curve cryptosystems. Math Comput 48:203–209
3. Downey F, Leong B, Seith R (1981) Computing sequences with addition chains. SIAM J Comput 10:638–646
4. Noma AM, Muhammed A, Mohamed MA, Zulkarnain ZA (2017) A review on heuristics for addition chain problem: towards efficient public key cryptosystems. J Comput Sci 13(8):275–289
5. Cruz-Cortés N, Rodríguez-Henríquez F, Juárez-Morales R, Coello Coello CA (2005) Finding optimal addition chains using a genetic algorithm approach. In: Hao Y et al (eds) Computational intelligence and security (CIS 2005). LNCS, vol 3801. Springer, Berlin, Heidelberg
6. Osorio-Hernandez L, Mezura-Montes E, Cruz-Cortes N, Rodriguez-Henriquez F (2009) A genetic algorithm with repair and local search mechanisms able to find minimal length addition chains for small exponents. In: IEEE congress on evolutionary computation (CEC 2009), pp 1422–1429
7. Rodriguez-Cristerna A, Torres-Jimenez J (2013) A genetic algorithm for the problem of minimal Brauer chains for large exponents. In: Melin P, Castillo O (eds) Soft computing applications in optimization, control, and recognition. Studies in fuzziness and soft computing, vol 294. Springer, Berlin, Heidelberg
8. Nedjah N, Mourelle LDM (2006) Towards minimal addition chains using ant colony optimization. J. Math Model Algorithms 5:525–543
9. Cruz-Cortés N, Rodríguez-Henríquez F, JuárezMorales R, Coello-Coello CA (2008) An artificial immune system heuristic for generating short addition chains. IEEE Trans Evol Comput 12:1–24
10. León-Javier A, Cruz-Cortés N, Moreno-Armendáriz MA, Orantes-Jiménez S (2009) Finding minimal addition chains with a particle swarm optimization algorithm. In: Aguirre AH, Borja RM, Garciá CAR (eds) Advances in artificial intelligence (MICAI 2009). LNCS, vol 5845. Springer, Berlin, Heidelberg
11. Jose-Garcia A, Romero-Monsivais H, Hernandez-Morales CG, Rodriguez-Cristerna A, Rivera-Islas I, Torres-Jimenez J (2011) A simulated annealing algorithm for the problem of minimal addition chains. In: Antunes L, Pinto HS (eds) Progress in artificial intelligence (EPIA 2011). LNCS, vol 7026. Springer, Berlin, Heidelberg (2011)
12. Dominguez-Isidro S, Mezura-Montes E, Cruz-Cortés N, Rodríguez-Henríquez F (2015) Evolutionary programming for the length minimization of addition chains. Eng Appl Artif Intell 37:125–134
13. Knuth DE (1981) The art of computer programming. Seminumeral algorithms, vol 2, 2nd edn. Addison-Wesley
14. Okeya K, Schmidt-Samoa K, Spahn C, Takagi T (2004) Signed binary representations revisited. In: Proceedings of CRYPTO'2004. LNCS 3152, pp 123–139
15. Balasubramaniam P, Karthikeyan E (2007) Elliptic curve scalar multiplication algorithm using complementary recoding. Appl Math Comput 190:51–56
16. Mohamed MA, Md Said MR, Mohd Atan KA, Ahmad Zulkarnain Z (2011) Shorter addition chain for smooth integers using decomposition method. Int J Comput Math 88(11):2222–2232
17. Mohamed MA, Mohd Atan KA (2012) Rule based representation of integer for a new addition chain method. Appl Math Sci 6(30):1497–1503
18. Mohamed MA, Said MRMd (2015) A hybrid addition chain method for faster scalar multiplication. Wseas Trans Commun 14:144–152
19. Mohamed MA, Ahmad A, Mohamed RR, Said MRM (2017) Shorter addition-subtraction chain with signed composition method. Int J Eng Technol 9(2):299–308

Integration of Iris Biometrics in Automated Teller Machines for Enhanced User Authentication

Kennedy Okokpujie, Etinosa Noma-Osaghae, Olatunji Okesola, Osemwegie Omoruyi, Chinonso Okereke, Samuel John and Imhade P. Okokpujie

Abstract The ubiquitous Automatic Teller Machine that revolutionized the way monetary transactions are carried out the world over is currently riddled with several security challenges. Top on the list of these challenges are the thefts and frauds associated with the ever popular Personal Identification Number based automatic teller machines. A lot of suggestions and proposals have been made in recent times, on how to combat the menace of automatic teller machine frauds. Biometrics is one of the most promising tools that have the capacity to put the nefarious activities around automatic teller machines in check. This paper proposes a cheap and economic iris biometric based automatic teller machine, built around a microcontroller, iris scanner and a robust database. The designed and implemented prototype is capable of checkmating automatic teller machine fraud and it is also easy to implement in developing nations.

Keywords Biometric · Automatic teller machine · Iris recognition
Authentication

1 General Introduction

There was a time when financial institutions like banks did their transactions manually. This made attending to customers a tedious and sometimes, a frustrating task [1]. Long queues in banks with the attendant waste of time and energy was the order of the day. Lots of people go to banks and may not be able to accomplish what they went to do in the bank [2].

K. Okokpujie (✉) · E. Noma-Osaghae · O. Okesola · O. Omoruyi · C. Okereke
S. John · I. P. Okokpujie
Department of Electrical and Information Engineering, Covenant University,
Ota, Ogun State, Nigeria
e-mail: kennedy.okokpujie@covenantuniversity.edu.ng

E. Noma-Osaghae
e-mail: etinosa.noma-osaghae@covenantuniversity.edu.ng

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_23

Today, a device called the Automatic Teller Machine (ATM) has made banking transactions easier. The ATM as the name implies can carry out common banking operations like deposits, withdrawals, transfers and payments without the need of any intervening human agent [3, 4]. The revolutionary machine has roots that are deep-seated in all facets of a bank's customer records. This singular fact gives the ATM its famed reputation as the "customer's cashier". The machine quickly gained wide acceptance because of its unique ability to provide 24 h of service every day [5].

The ATM's wide and burgeoning acceptance has led to some security concerns due to the nefarious activities of some unscrupulous elements called fraudsters. The most popular line of protection for most ATM machines today is the Personal Identification Number (PIN). This line of protection is fragile and can be lost or forgotten. In the event of an ATM card theft, the PIN could be guessed correctly by trained hackers. This has given rise to a whole lot of security concerns especially on the part of the customers who use ATMs. The call for a better, more secure and reliable means of authentication when using ATMs has risen in recent times. Customers and stakeholders the world over are demanding that ATMs be made more secure to use for financial transactions. The demand for a more sophisticated ATM authentication process is louder now than ever [6].

Biometrics shows a lot of promise [7]. The unique and immutable characteristics got from biometric identifiers like fingerprints, iris, face, voice and gait has been explored as security layers for several applications especially in law enforcement agencies and private security outfits. Biometrics can provide an additional layer of protection or authentication for ATM transactions [8]. The use of biometrics is so promising because it is very rare to have two persons with exactly the same biometric traits [9].

The present means of authenticating ATM transactions cannot differentiate between impostors and genuine users. The PIN codes used by customers could be acquired fraudulently by criminals. Criminals have become so advanced in their tactics that they use secret scanners to acquire information about the PINs and card used during ATM transactions. ATM fraud can be ATM transaction fraud [10]. Proactive steps must be taken by financial institutions to ensure that customers have absolute trust in their ability to shield them from ATM transaction fraud.

This paper proposes the incorporation of iris biometrics into ATM terminals. The iris biometric ATM terminal prototype adds an additional layer of protection to the conventional PIN code ATM terminal to strengthen the authentication process for conventional and niche ATM transactions. The proposed biometric ATM uses two means of authentication, namely, PIN code and iris. At the time the customer opens an account with the bank, the iris' template, along with the user particulars are stored in a database. After enrolment, a default PIN is issued to the user. To get authenticated for a transaction, the user keys in the PIN code. If the PIN is correct, the proposed system prompts the user to go through the motions of getting their iris scanned by the iris scanner and compared with the iris template of the user stored in the database. If there is a match, the user is allowed to carry out the transaction. The process from keying in the PIN to verifying the iris biometric details is repeated for each new transaction.

The proposed system can reduce ATM fraud drastically. The cost of iris scanners, variation in heights, lighting and motion tends to limit the way the iris can be applied to an ATM. These notwithstanding, the proposed iris biometric ATM has a low false acceptance and rejection rate.

2 Related Works

Koteswari et al. emphasized the security lapses in the use of just PIN codes to verify users during ATM transactions. Among the security lapses highlighted in their work were impersonation, theft, the use of advanced technologies to guess PIN codes of stolen debit and credit cards. The researchers proposed the use of biometrics to checkmate the unpleasant security lapses of the PIN only ATMs. The authors suggested the fusion of fingerprint and iris biometric characteristics to authenticate users of ATMs. The researchers concluded that the use of biometrics to authenticate ATM transactions would drastically reduce fraudulent activities and make ATM transactions safer [11].

Okokpujie et al. carried out a research to see how an improved iris segmentation process can lead to better and more accurate recognition rates in iris recognition technology [12]. Their research involved the comparison of the integro-differential and circular Hough transforms approach to iris segmentation. The researchers strongly felt that the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) of any iris recognition technology can be greatly influenced by its iris segmentation process. The study tried to prove that spoofing iris recognition systems using artificial lens and paper prints can be made more difficult by integrating a more accurate segmentation process into an already efficient iris recognition algorithm. The study concluded that the circular Hough transform (Wilde's Model) offered a 1.67% better accuracy in comparison to the integro-differential segmentation model proposed by Okokpujie et al. [13].

Soares et al. proposed the use of fingerprint [14] and iris biometric controlled ATMs to checkmate the surge in fraud and theft at ATM terminals. The study completely replaced ATM cards and PINs with iris and fingerprint biometrics. The system first captures and matches the iris and fingerprint pattern of the user. After matching is done, a Global System for Communication module sends a one-time password (OTP) to the phone number registered by the user at the point of enrolment. The user is allowed to carry out transactions on the ATM after keying in the one-time password correctly into the ATM. The researchers collected and analyzed real-time data got form fingerprints and irises used for the experiment. The proposed system also had a blocking feature that blocks or prevents impostors and fraudsters from illegally accessing accounts [15].

Ahmad et al. proposed a card less ATM that uses biometrics and an Advanced Encryption Standard (AES) algorithm. The researchers detailed the specifications of the proposed system and stated that the system can be used for secure banking transactions that require small amounts of memory and high speed performance. The proposed card less ATM biometric machine had a throughput of about 19.016 Gbps [16].

Gatali et al. [17] carried out a qualitative study on adoption of biometric technologies in Canadian banking industry in comparison with other nations around the world. Their research looked at how well biometrics have been integrated into the operations of Canadian Banks, the specific biometric technologies influencing banking operations in Canada and the future of biometrics in Canada's banking sector. In all, the researchers felt that Canadian banks need to become open to more sophisticated levels of biometric integration.

Kassem proposed a unique ATM security system that uses multimodal biometrics. The system fused iris and fingerprint biometrics. The researchers used a minutiae matcher for fingerprints and hamming distance matcher for iris. The threshold was set at 0.6. The proposed system had a False Acceptance Rate of 0% and a run time of 32 s [18].

Yin et al. proposed and efficient iris image segmentation for ATM based on fuzzy entropy and graph cut. The researchers used reduced redundant computations in fuzzy entropy evaluation by employing an iterative calculation scheme. The probabilities of four (4) fuzzy events were used to define the cost of four (4) label assignments, namely, background, eyelash, pupil and iris. The researchers declared that when presented and segmented on a graph cut, the four (4) label assignments produced a result that was better than other existing segmentation techniques [19–21].

3 Methodology

The ATM's security is improved using iris recognition. The design is centered around the ATmega 128 chip. The iris camera and other peripherals are controlled by the embedded chip.

The functional block diagram is shown in Fig. 1.

The proposed iris biometric system consists of three phases, namely, Account Creation Phase, Biometric Enrolment Phase and Authentication Phase.

The authentication process is shown in Fig. 2:

The enrolment process is shown in Fig. 3.

The biometric enrolment process is monitored by the administrator and the aim of this phase is to tie the biometric trait (iris patterns) to the account number and PIN entered during the account creation stage. The details of which are shown in Fig. 3.

The algorithm of the software tasks for the enrolment process is given below:

```
Step 1: Initialize by pressing (button for enrolment)
Step 2: Key in Account number
Step 3: If account number is valid
Go to Step 4
Else Step 1
Step 4: Key in PIN
```

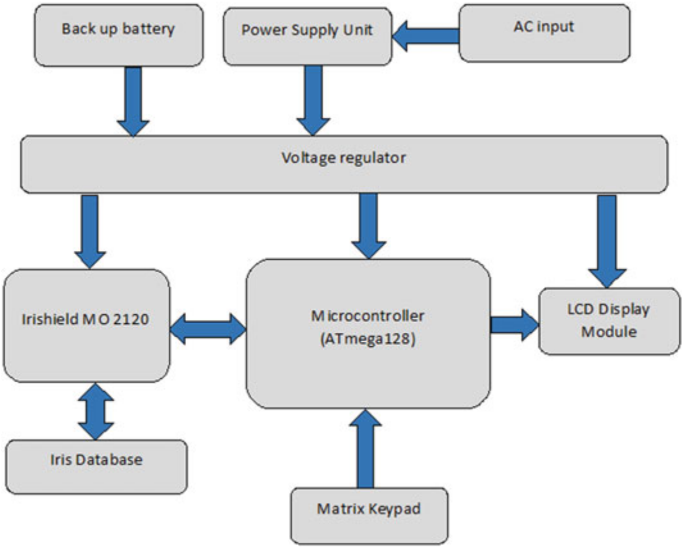


Fig. 1 Functional block diagram of the iris biometrics ATM system

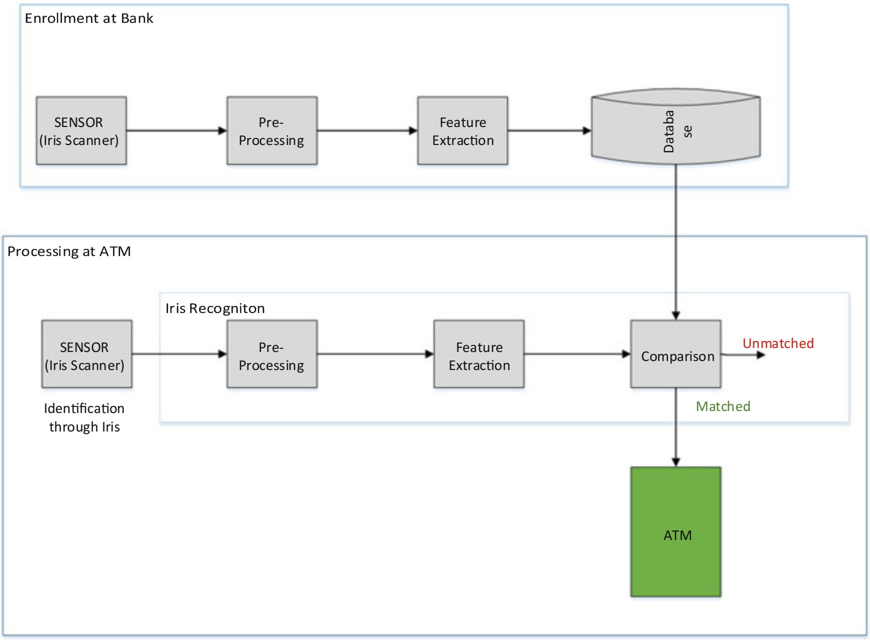


Fig. 2 Iris biometric ATM enrolment and authentication process

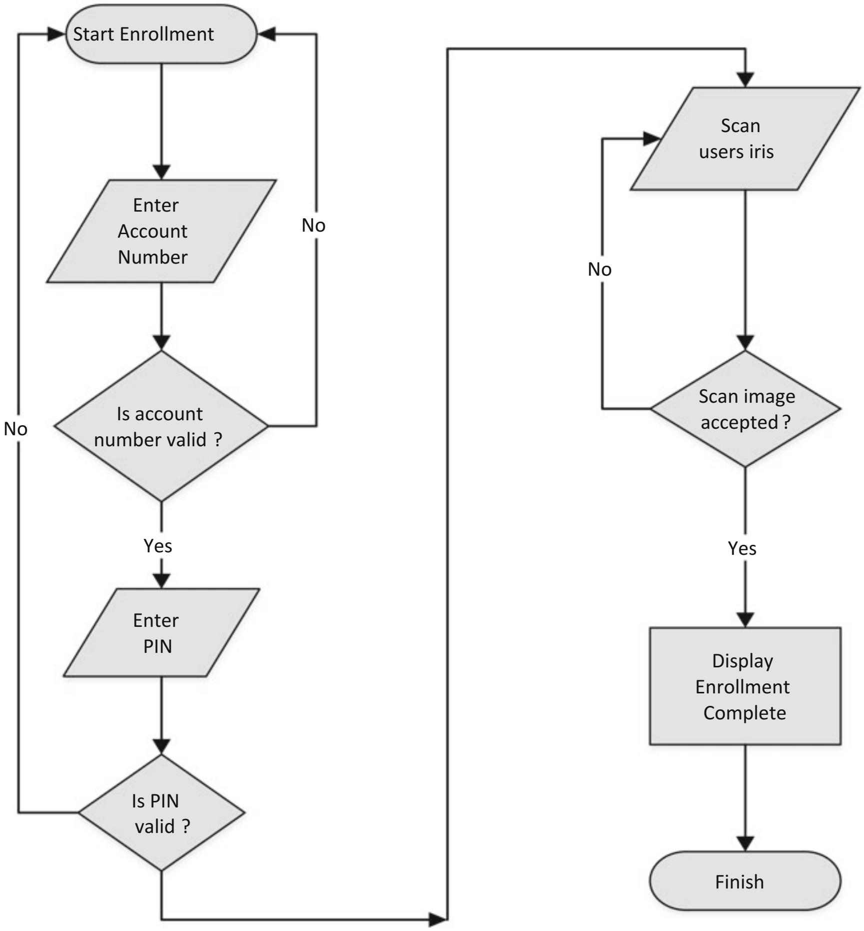


Fig. 3 The enrolment process

```
Step 5: If PIN is valid
Go to Step 6
Else
Go to Step 1
Step 6: Request for user’s iris for scan
Step 7: If scan is accepted
Go to Step 8
Else
Go to Step 6
Step 8: Display Enrolment Complete
Step 10: Press CANCEL to return to Home
```

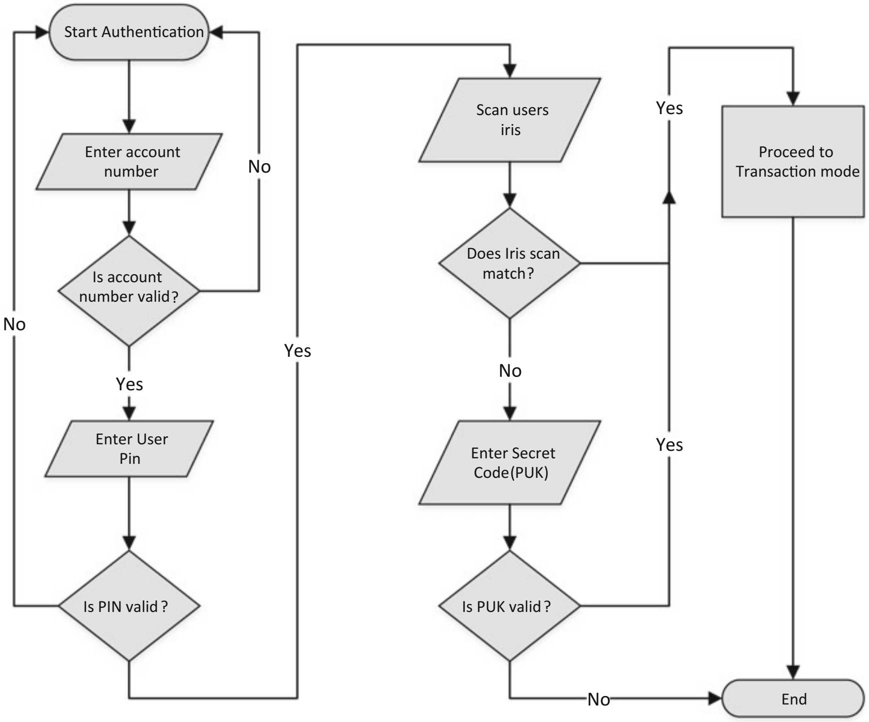


Fig. 4 The authentication process

The Authentication process is shown in Fig. 4. After the user iris has been scanned and enrolled with (necessary) data, the account number will be the user's ID. In order to get access into the ATM, the user has to key in an account number and PIN via the keypad, the system matches this information to an account and then authentication is done by matching the user's iris details with previously the enrolled one. The algorithm of the software task of the designed embedded system is:

```
Step 1: Initialize by pressing 'ENTER' keypad
Step 2: Key in account number
Step 3: If account number is valid
Go to Step 4
Else
Go to Step 1
Step 4: Key in user PIN
Step 5: If the PIN is valid
Go to Step 6
Else
Go to Step 1
```

```
Step 6: Request user's iris on iris scanner
Step 6: If the Iris pattern is matched
Go to Step 9
Else
Go to Step 7
Step 7: Request for Secret Code (PUK)
Step 8: If the Secret Code is valid
Go to Step 9
Else
Go to Step 1
Step 9: Transaction mode
```

4 Result Analysis

1. Sample acquisition errors measure environmental conditions surrounding the system and are quantified by means of:

   - Failure to Enroll Rate: The number of times the iris biometrics system rejected provided iris samples is zero.
   - Failure to Capture Rate: The number of times the iris biometrics system rejected provided iris samples is zero.

2. Recognition performance is a measure of how well the system is able to correctly match the biometric information from the same person and avoid false matches of biometric information from different people. The measures used to quantify biometric accuracy of a biometric system are:

   - True Match Rate (TMR): a 100% match was confirmed from tests carried out with the iris biometric system, as users were correctly matched.
   - False Match Rate (FMR): from tests carried out with the system, there was no false match, as access wasn't granted to wrong users. (FMR = 0%)
     $\mathrm{FMR} = \mathrm{NFA} \div \mathrm{NIIA} = 0 \div 15 = 0$
     NFA—number of false acceptance
     NIIA—number of impostor identification attempts = 15
   - False Non-Match Rate (FNMR): from tests carried out with the system, there was no false rejection, as access was granted to all authorized users (FNMR = 0).
     $\mathrm{FNMR} = \mathrm{NFR} \div \mathrm{NEIA} = 0 \div 30 = 0$
     NFR—number of false rejection
     NEIA—number of enrollee identification attempt.

5 Conclusion

An ATM prototype that uses iris biometric system for a more reliable authentication of bank customers at ATM terminals has been designed and implemented. The proposed system has overcome the present vulnerabilities with current means of authentication and guarantees a secure way of carrying out transactions at ATM terminals. The designed system is capable of reducing drastically the rate of ATM fraud and restoring customer trust in the banking system.

The designed system:

- Introduces iris biometrics as the second level of authentication which helps to improve ATM security.
- Improves upon the present one level authentication process (i.e. the use of card and PIN only) which is vulnerable to fraud.
- The present ATM machines can be upgraded to biometric ATMs thereby reducing the cost of purchase of new ones.
- Its implementation will bring back customer confidence in the use of ATM and in the banking system.

Acknowledgements This paper is sponsored by Covenant University, Ota, Ogun State, Nigeria.

Further Work An enhanced multimodal biometric ATM would be designed and implemented.

References

1. Okokpujie K, Noma-Osaghae E, John S, Jumbo PC (2017) Automatic home appliance switching using speech recognition software and embedded system. In: International conference on computing networking and informatics (ICCNI), 2017, pp 1–4
2. Okokpujie KO, Uduehi OO, Edeko FO (2016) An innovative technique in ATM security: an enhanced biometric ATM with GSM feedback mechanism. J Electr Electron Eng (JEEE), vol 12, pp 68–81
3. Wang Y, Zhang Y, Sheu PC-Y, Li X, Guo H (2012) The formal design model of an automatic teller machine (ATM). In: Breakthroughs in software science and computational intelligence. IGI Global, pp 263–287
4. Okokpujie K, Olajide F, John S, Kennedy CG (2016) Implementation of the enhanced fingerprint authentication in the ATM system using ATmega128. In: Proceedings of the international conference on security and management (SAM), 2016, p 258
5. Okokpujie K, Uduehi O, Edeko F (2015) An enhanced biometric ATM with GSM feedback mechanism. J Electr Electron Eng 12:68–81
6. Atuegwu C, Okokpujie KO, Noma-Osaghae E (2017) A bimodal biometric student attendance system
7. Majekodunmi TO, Idachaba FE (2011) A review of the fingerprint, speaker recognition, face recognition and iris recognition based biometric identification technologies
8. Okokpujie K, Noma-Osaghae E, John S, Oputa R (2017) Development of a facial recognition system with email identification message relay mechanism. In: 2017 international conference on computing networking and informatics (ICCNI), 2017, pp 1–6

9. Daramola SA, Adefuminiyi MA, John TM (2016) Review and proposed methodology for a lecture attendance system using neural network. In: Proceedings of the world congress on engineering
10. John S, Anele C, Kennedy OO, Olajide F, Kennedy CG (2016) Realtime fraud detection in the banking sector using data mining techniques/algorithm. In: 2016 international conference on computational science and computational intelligence (CSCI), 2016, pp 1186–1191
11. Koteswari S, Paul PJ (2017) A survey: fusion of fingerprint and iris for ATM services
12. Badejo JA, Atayero AA, Ibiyemi TS (2016) A robust preprocessing algorithm for iris segmentation from low contrast eye images. In: Future technologies conference (FTC), 2016, pp 567–576
13. Okokpujie K, Noma-Osaghae E, John S, Ajulibe A (2017) An improved iris segmentation technique using circular Hough transform. In: International conference on information theoretic security, 2017, pp 203–211
14. Okokpujie K, Etinosa N-O, John S, Joy E (2017) Comparative analysis of fingerprint preprocessing algorithms for electronic voting processes. In: International conference on information theoretic security, 2017, pp 212–219
15. Soares J, Gaikwad A (2016) Fingerprint and iris biometric controlled smart banking machine embedded with GSM technology for OTP. In: Automatic control and dynamic optimization techniques (ICACDOT), 2016, pp 409–414
16. Ahmad N, Rifen AAM, Wahab MHA (2016) AES cardless automatic teller machine (ATM) biometric security system design using FPGA implementation. In: IOP conference series: materials science and engineering, 2016, p 012113
17. Gatali IF, Lee KY, Park SU, Kang J (2016) A qualitative study on adoption of biometrics technologies: Canadian banking industry. In: Proceedings of the 18th annual international conference on electronic commerce: e-Commerce in smart connected world, 2016, p 20
18. Kassem MA, Mekky NE, EL-Awady RM (2014) An enhanced ATM security system using multimodal biometric strategy. Int J Electr Comput Sci (IJECS-IJENS) 14:9–16
19. Yin S, Zhao X, Wang W, Gong M (2014) Efficient multilevel image segmentation through fuzzy entropy maximization and graph cut optimization. Pattern Recognit 47:2894–2907
20. Okokpujie KO, Etinosa N-O, Okesola OJ, Samuel JN, Robert O (2017) Design and implementation of a student attendance system using iris biometric recognition. In: Computational science and computational intelligence (CSCI), 2017, Las Vegas, USA
21. Etinosa N-O, Okereke C, Robert O, Okesola OJ, Okokpujie KO (2017) Design and implementation of an iris biometric door access control system. In: Computational science and computational intelligence (CSCI), 2017, Las Vegas, USA

Fingerprint Biometric Authentication Based Point of Sale Terminal

Kennedy Okokpujie, Etinosa Noma-Osaghae, Olatunji Okesola, Osemwegie Omoruyi, Chinonso Okereke, Samuel John and Imhade P. Okokpujie

Abstract Retail businesses that are not transacted online still represent a substantial amount of retail deals that are closed on a daily basis. Retail business owners and customers continue to explore other means of ensuring payments made with Point of Sale (POS) devices are done securely. This paper proposes the incorporation of fingerprint biometric recognition as an additional layer of protection to the customary pin and password requirements to gain permission to pay for goods purchased and services rendered using point of sale devices. The proposed fingerprint biometric recognition point of sale device has zero false match and false non-match rate. This strengthens the present authentication process that makes use of pins and passwords that are prone to fraud and solidifies the trust and confidence users place on point of sale devices.

Keywords Biometric · Point of sale · Fingerprint · Recognition · Authentication

1 General Introduction

Security continues to be a matter of great concern to individuals, organizations and governments. Shopping malls, restaurants and other commercial centers with very large number of customers are willing to deploy new and more effective means of checkmating fraud and impersonation. This paper explores the possibility of adding another layer of protection at point of sale terminals [1].

K. Okokpujie (✉) · E. Noma-Osaghae · O. Okesola · O. Omoruyi · C. Okereke
S. John · I. P. Okokpujie
Department of Electrical and Information Engineering,
Covenant University, Ota, Ogun State, Nigeria
e-mail: kennedy.okokpujie@covenantuniversity.edu.ng

E. Noma-Osaghae
e-mail: etinosa.noma-osaghae@covenantuniversity.edu.ng

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_24

Fingerprints are one of the several biological, physiological and behavioral traits used in biometric recognition systems [2]. The fingerprint is unique to each individual and is not the same even in identical twins. Injuries only temporarily mar the fingerprint's pattern. When injuries to the finger become healed, the original pattern of the fingerprint is restored [3, 4].

Fingerprint recognition systems usually involve two phases, namely, enrolment and authentication (or verification) [5]. During enrolment, the image of the fingerprint is captured by a special device. The captured image is converted into digital form and its minutia points extracted by an algorithm or special sensor [6]. The extracted minutia points are used to form a template that can be stored alongside the user's particulars in a database [7]. The extracted minutia forms the basis upon which every individual can be uniquely identified [8].

During authentication, the query print can be compared with just one enrolled print to prevent multiple individuals from using one identity [9]. This is done in a process called verification (1:1 matching). But it is called Identification (1:N matching)—N is the number of individuals which have been successfully enrolled, when the query print is compared with all enrolled prints in the database to determine if an individual is known under a different name (or duplicate) [10, 11].

Aging and manual labor that tend to flatten the friction ridges [12, 13], injuries that can distort fingerprint patterns [14, 15], capture area of the sensor and the resolution of the captured fingerprint image can adversely affect the ability of a fingerprint recognition system to correctly authenticate users.

In recent times, more commercial centers have implemented one form of biometric recognition or the other in some of their daily processes [13, 16] and as the technology supporting biometric recognition becomes more ubiquitous and cheaper, there are bound to be more areas of application for biometric technology in commercial centers such as shopping malls and parks.

Retail entrepreneurs are consistently devising new ways of defrauding users of point of sale (POS) terminals. Thus this paper proposes the need to add an additional layer of authentication to POS devices. The proposal being the compulsory requirement of biometric authentication (fingerprint) for every payment made via POS devices.

This paper vividly describes the design and implementation of a prototype POS device that uses fingerprint biometric for payment confirmation. The main objective was to develop a prototype point of sale device that uses fingerprint to authenticate payments for goods purchased or services rendered.

The proposed system makes use of embedded systems and fingerprint biometric technologies. Each user will be given an RFID (Radio Frequency Identification) card with which they can access the point of sale device. During enrolment, a fingerprint sensor captures the user's fingerprint(s), an algorithm extracts the minutia points from the digitized fingerprint image, and a template of the extracted minutia points is made and stored with the user's particulars in a database. During authentication, the RFID card is placed in the slot provided and the first layer of protection, which is the PIN (Personal Identification Number), is keyed into the point of sale device. After pin verification, the device initiates the second layer of protection by carrying out fingerprint verification. If verified, the user is granted permission to carry out transactions using the point of sale device. The whole process is repeated for every new transaction. Incorrect placement of the user's finger on the fingerprint sensor in terms of pressure and position, can adversely affect the performance of the proposed system.

2 Related Works

Mansfield-Devine stressed the inevitability of biometrics in retail businesses. The author acknowledged the fact that biometrics have found acceptance in diverse fields of applications where personal identification is paramount. Some of these areas include, voting and access to secure premises. Steve also noted the concerns that will arise from implementing biometric payment authentication in retail businesses [17]. He gave a glimpse into the training that would be required for both customers and vendors if biometric payment authentication is rolled out in large scale.

Ghosh et al. proposed a system of carrying out cashless transfer of fund between cards using near-field communication. The authors also designed and implemented a payment system that uses fingerprint biometrics to authenticate the payments made by customers. The designed system could also make use of the Global System for Mobile Communication (GSM) to effect payments [18].

Masalkar et al. proposed a 2D barcode based mobile payment system that can be implemented on a Point of Sale device [19]. A secured biometric payment model that uses tokenization and fingerprint biometrics was proposed by Garg and Garg [20]. The authors declared that the designed system could be used to make payments at a merchant point of sale terminal without the need of cards. Alimi et al. proposed a Signature Dynamics based user authentication system that turns mobile devices into point of sale terminals [21–23].

3 Methodology

The designed system uses ATmega128 as the embedded system linked to the fingerprint recognition technology. The user's minutiae point was used for authentication (Fig. 1).

The fingerprint scanner was used to scan the finger. Minutiae points were extracted from the acquired fingerprint image and used to form templates that can be stored during enrolment and matched during verification. The fingerprint scanner used to implement the designed fingerprint biometric point of sale device had a high speed fingerprint algorithm engine that could effectively carry out 1:N identification, 1:1 verification, fingerprint feature extraction, data read/write functions and security level setting.

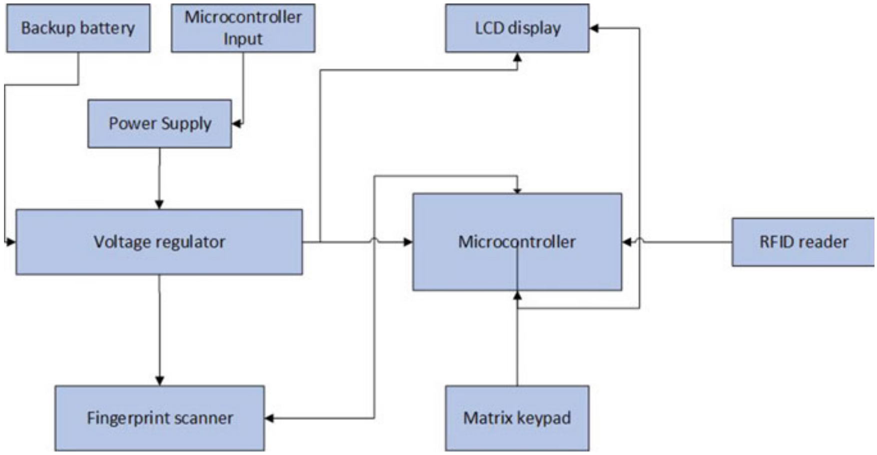


Fig. 1 Block diagram of the fingerprint biometrics POS system

The fingerprint scanning device had the following specifications:

1. Template size of 512 Bytes.
2. Security Level of 1–5 with 3 as the default.
3. Fingerprint reader module size of 20 mm by 32 mm.
4. Effective collection area of 11 mm by 15 mm.
5. Scanner resolution of 508 DPI.
6. Fingerprint pre-treatment time response of less than 0.25.

The ATmega microcontroller has the following specifications:

7. A bus size of 8-bits.
8. 4 kb of RAM.
9. A processing speed of up to 16 MHz.

RFID cards were used in place of ATM cards and thus an RFID card reader was part to the overall system design. During verification, RFID card reader sends information to the microcontroller that is used to identify user accounts for which fingerprint biometric recognition must be carried out.

The data bus of the alphanumeric liquid crystal display was used to create a visual interface that was used as the primary means of monitoring what the implemented system does per time.

A multiplexed matrix keypad was used as the input device. The eight wire cable energy from the keypad was wired into the microcontroller to enable users communicate with the microcontroller. Each key on the keyboard creates a scan code that can be remapped to specific functions in the software.

The C language was used to create the program that ran on the implemented system. To this end, a compiler with an Integrated Development Environment (IDE) was used to write, compile and debug the program. The major states through which the implemented system can pass include:

- Idle.
- RFID cards and PIN request.
- Biometric Authentication.
- PUK request.
- Authenticated.
- Balance.
- Transaction.
- Change Pin.

The system's fingerprint biometric recognition system makes use of a matching score to set the threshold beyond which a user is granted permission to make transactions with the POS device and below which the user is denied access. This match score is given whenever the "query print" is compared with the "stored print" for pairwise similarities.

3.1 Enrolment

In this phase, the Fingerprint reader was used to identify or verify a user. The main objective of this enrolment phase is to create a profile for the user in the database (Fig. 2).

The enrolment process, done by the administrator, gets the user details, fingerprint and saves the information.

The algorithm of the software tasks for the enrolment process is given below:

```
STEP 1: Initialize by pressing (button for enrolment)
STEP 2: Enter Account number
STEP 3: If account number is valid
GOTO STEP 4
ELSE STEP 1
STEP 4: Enter PIN
STEP 5: If PIN is valid
GOTO STEP 6
ELSE
GOTO STEP 1
STEP 6: Request for user's iris for scan
STEP 7: If scan is accepted
GOTO STEP 8
ELSE
GOTO STEP 6
STEP 8: Display Enrolment Complete
STEP 10: Press CANCEL to return to Home
```

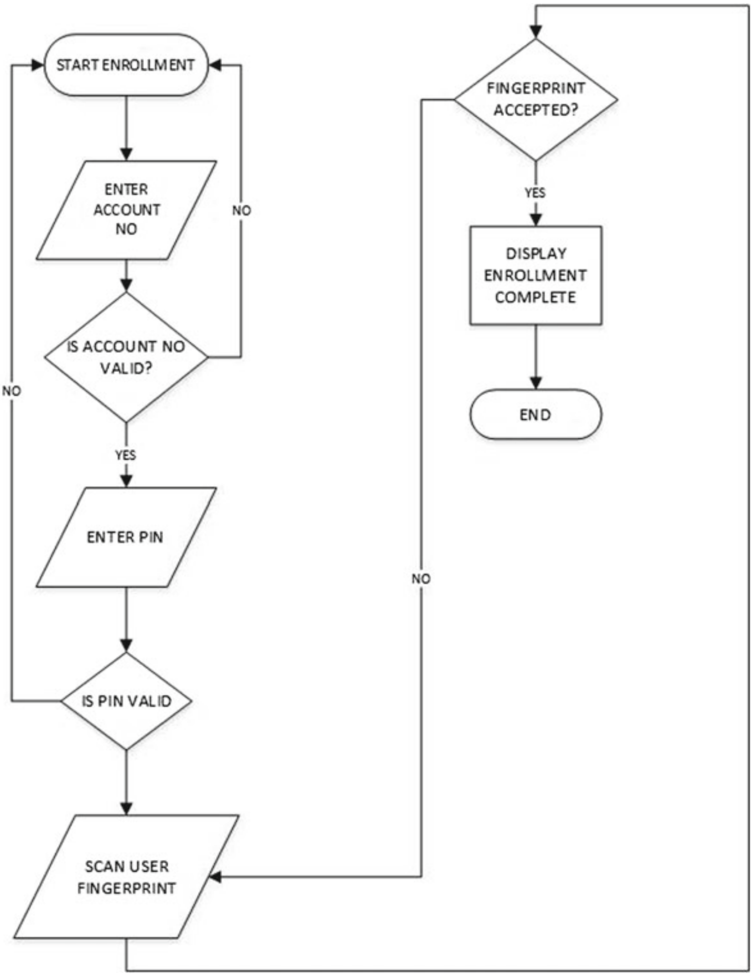


Fig. 2 Flowchart for the enrolment process

3.2 Authentication

Once an individual has been enrolled into the system, the user swipes the RFID card against the reader module, keys in the account pass and then follows the process for feature extraction. Verification implies a one-to-one match requiring the user to provide the RFID card and account pin as a means of identification. The biometric sample along with the provided identification is compared to the previously stored information. If there is a match with the fingerprint pattern enrolled, access is provided to the account, and otherwise it is rejected. This is described by the flowchart in Fig. 3.

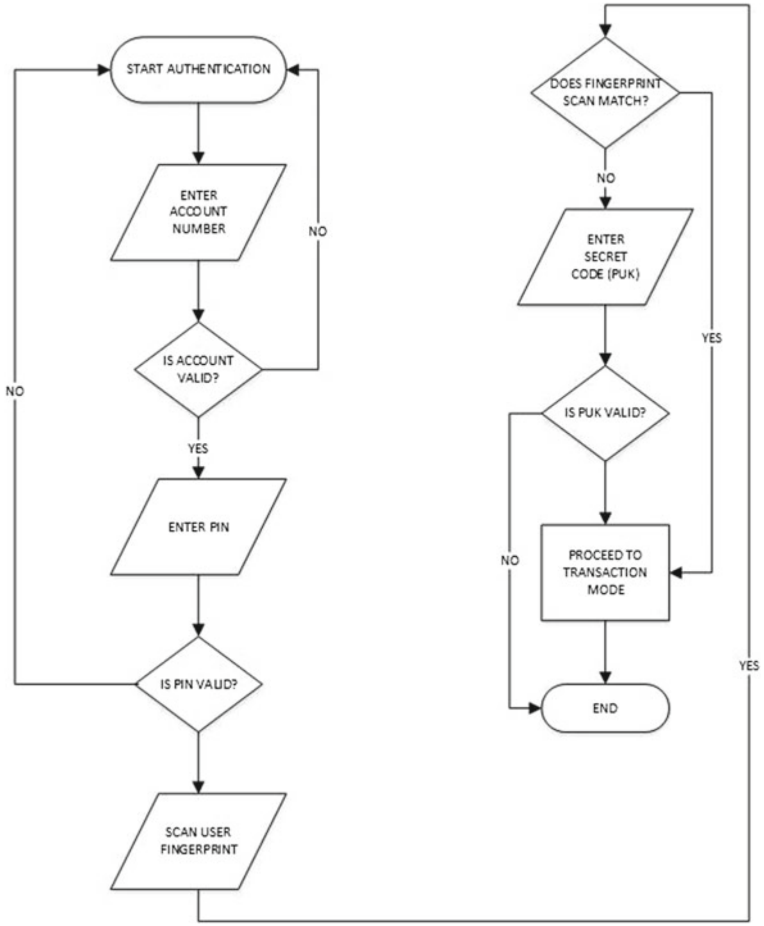


Fig. 3 Flowchart for payment process

In order to get access into the POS after enrolment, the user slots in the RFID card into the device and keys in the account pin via the keypad. The system matches this number to an account and then does biometric verification. The algorithm of the software task of the designed embedded system is:

```
STEP 1: Initialize by pressing 'ENTER' keypad
STEP 2: Enter account number
STEP 3: If account number is valid
STEP 6: Request user's iris (as input) on iris scanner
STEP 6: If the Iris pattern is matched
GOTO STEP 9
ELSE
GOTO STEP 7
GOTO STEP 4
ELSE
GOTO STEP 1
STEP 4: Enter user PIN
STEP 5: If the PIN is valid
GOTO STEP6
ELSE
GOTO STEP 1
STEP 7: Request for Secret Code (PUK)
STEP 8: If the Secret Code is valid
GOTO STEP 9
ELSE
GOTO STEP 1
STEP 9: Transaction mode
```

4 Conclusion

The performance of the implemented POS device was gauged using the False Non-Matching Rate (FNMR) and the False Matching Rate. The False Non-Matching rate is the frequency with which the fingerprint biometric recognition system declares fingerprints from the same person as different or a non-match. This did not occur at all and thus the implemented system has a zero (0) False Non-Match Rate.

The False Matching Rate (FMR) is the frequency with which the fingerprint biometric recognition system declares fingerprints from different persons as the same or a match. This did not also occur at all and thus the implemented system has a zero (0) false matching rate. Upon testing other functionalities after access is granted by the fingerprint biometric system, it was discovered that the POS device worked perfectly well. The device accurately paid for goods bought and services rendered.

Acknowledgements This paper was sponsored by Covenant University, Ota, Ogun State, Nigeria.

Further Work A conscious and active incorporation of a liveness detector into the designed and implemented fingerprint biometric POS device.

References

1. Okokpujie K, Uduehi O, Edeko F (2015) An enhanced biometric ATM with GSM feedback mechanism. J Electr Electron Eng 12:68–81
2. Okokpujie K, Noma-Osaghae E, John S, Ajulibe A (2017) An improved iris segmentation technique using circular Hough transform. In: International conference on information theoretic security, 2017, pp 203–211

3. Kaur R, Sandhu PS, Kamra A (2010) A novel method for fingerprint feature extraction. In: 2010 international conference on networking and information technology (ICNIT), 2010, pp 1–5
4. Badejo JA, Atayero AA, Ibiyemi TS (2016) A robust preprocessing algorithm for iris segmentation from low contrast eye images. In: Future technologies conference (FTC), 2016, pp 567–576
5. Atuegwu C, Okokpujie KO, Noma-Osaghae E (2017) A bimodal biometric student attendance system
6. Okokpujie K, Noma-Osaghae E, John S, Jumbo PC (2017) Automatic home appliance switching using speech recognition software and embedded system. In: 2017 international conference on computing networking and informatics (ICCNI), 2017, pp 1–4
7. Okokpujie K, Noma-Osaghae E, John S, Oputa R (2017) Development of a facial recognition system with email identification message relay mechanism. In: 2017 international conference on computing networking and informatics (ICCNI), 2017, pp 1–6
8. Tukur A (2015) Fingerprint recognition and matching using Matlab. Int J Eng Sci (IJES) 4:01–06
9. Okokpujie KO, Uduehi OO, Edeko FO (2016) An innovative technique in ATM security: an enhanced biometric ATM with GSM feedback mechanism. J Electr Electron Eng (JEEE), vol 12, pp 68–81
10. Barham ZS, Mousa A (2011) Fingerprint recognition using MATLAB. Bachelor's Dissertation, vol 5, p 17
11. Daramola SA, Adefuminiyi MA, John TM (2016) Review and proposed methodology for a lecture attendance system using neural network. In: Proceedings of the world congress on engineering
12. Jain AK, Feng J, Nandakumar K (2010) Fingerprint matching. Computer 43
13. Okokpujie K, Olajide F, John S, Kennedy CG (2016) Implementation of the enhanced fingerprint authentication in the ATM system using ATmega128. In: Proceedings of the international conference on security and management (SAM), 2016, p 258
14. Dhundhwal P, Maan N (2014) Design and implementation of enhancement feature extraction and matching of a fingerprint image. Int J Eng Trends Technol 13:184–190
15. Okokpujie K, Etinosa N-O, John S, Joy E (2017) Comparative analysis of fingerprint preprocessing algorithms for electronic voting processes. In: International conference on information theoretic security, 2017, pp 212–219
16. Majekodunmi TO, Idachaba FE (2011) A review of the fingerprint, speaker recognition, face recognition and iris recognition based biometric identification technologies
17. Mansfield-Devine S (2013) Biometrics in retail. Biom Technol Today 2013:5–8
18. Ghosh S, Majumder A, Goswami J, Kumar A, Mohanty SP, Bhattacharyya BK (2017) Swing-Pay: one card meets all user payment and identity needs: a digital card module using nfc and biometric authentication for peer-to-peer payment. IEEE Consum Electron Mag 6:82–93
19. Masalkar PA, Singh U, Shinde S (2015) 2D barcode based mobile payment system with biometric security. Transportation 2
20. Garg RK, Garg N (2015) Developing secured biometric payments model using tokenization. In: 2015 international conference on soft computing techniques and implementations (ICSCTI), 2015, pp 110–112
21. Alimi V, Rosenberger C, Vernois S (2013) A mobile contactless point of sale enhanced by the NFC and biometric technologies. Int J Internet Technol Secur Trans 5:1–17
22. Etinosa N-O, Okereke C, Robert O, Okesola OJ, Okokpujie KO (2017) Design and implementation of an iris biometric door access control system. In: Computational science and computational intelligence (CSCI), 2017, Las Vegas, USA
23. Okokpujie KO, Etinosa N-O, Okesola OJ, Samuel JN, Robert O (2017) Design and implementation of a student attendance system using iris biometric recognition. In: Computational science and computational intelligence (CSCI), 2017, Las Vegas, USA

A User Study: Abuse Cases Derived from Use Case Description and CAPEC Attack Patterns

Imano Williams and Xiaohong Yuan

Abstract Nowadays, developers should incorporate software security best practices from the early stages of the software development lifecycle to build more robust software against software security attacks. However, incorporating security practices at the early stages of the SDLC is difficult for novice software developers that do not have a systematic approach to address security issues. In this paper, we proposed a preliminary method to derive abuse cases, one of software security best practices, based on use case description and attack patterns and then evaluate the method in a user study. We investigated the effectiveness of the proposed method to help novices develop abuse cases and gained insights on how a novice of software security would select keywords from use case descriptions, and select relevant attack patterns for developing abuse cases. Our main findings were (1) the approaches participants used to select the keywords and the attack patterns as they related to the use cases; (2) the approach used to select relevant attack patterns; (3) the relationship between the keywords and the attack patterns; and (4) use case based on the textual content showed the method can be effective in assisting non-experts to create abuse cases. Finally, we suggest possible approaches to select keywords more effectively and the implication of using an inference engine to build relationships between use cases and attack patterns.

Keywords Software security · Attack patterns · Abuse cases

I. Williams (✉) · X. Yuan
Department of Computer Science, North Carolina A&T State University,
Greensboro, NC 27411, USA
e-mail: irwilli1@aggies.ncat.edu

X. Yuan
e-mail: xhyuan@ncat.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_25

1 Introduction

Nowadays, secure software development practices should be incorporated from the early stages of the Software Development Life Cycle (SDLC). However, incorporating security practices at the early stages of the SDLC maybe difficult for non-expert or novice software developers if there is no systematic approach. Creating abuse cases are one of the several security practices that can be used in the early stages of the SDLC. Use cases are used to describe the interaction between the intended user and the software application. On the other hand, abuse cases describe the interaction between a malicious user and the software application that may cause unexpected system behavior or exploit vulnerabilities [1, 2]. Abuse cases can be created based on a set of requirements and use cases, and a list of attack patterns [3]. Additionally, informed brainstorming of the system to be built has been suggested for creating abuse cases [4]. However, software security experts are more capable of creating meaningful and useful abuse cases than non-experts using the brainstorming approach. Therefore, we proposed a method for deriving abuse cases for a use case based on the keywords from use case description and the attack patterns in Common Attack Pattern Enumeration Classification (CAPEC) [5].

The main purpose of this study is to conduct a user study on the proposed method with the following objectives: (1) evaluate the effectiveness of the proposed method to help non-experts develop abuse cases; and (2) gain insights on how a non-expert of software security would select keywords from use case descriptions, and select relevant attack patterns for developing abuse cases.

The major contribution of our study is: (1) Demonstrate how non-experts can strategize an approach to retrieve relevant attack patterns based on selected keywords that may map use cases to attack patterns and (2) Use the results of the study to help understand the approaches that can be used to create abuse cases based on textual description of use cases.

The rest of the paper is organized as follows: Sect. 2 reviews related work. Section 3 presents the proposed method. Section 4 describes the user study of the proposed method. Section 5 discusses the user study results. Finally, Sect. 6 concludes the paper.

2 Related Work

Sindre and Opdahl [6] proposed misuse or abuse cases as a method for security requirements elicitation. They used misuse cases to address potential threats during system design and provided supplementary functionalities to help mitigate the threats. Alexander [7] used misuse cases in a practical environment to address security and safety issues in the trade-offs between different software designs. Both works used abuse cases in the design phase to address security issues. However, the method we propose uses abuse cases to capture security issues during the requirements phase, and focuses on how to create abuse cases based on use cases.

A Hierarchy-Driven Approach was proposed by Pauli and Engebretson [8] that uses attack patterns from CAPEC to teach students how to semi-formally represent an attacker's perspectives. Kaiya et al. [9] proposed a method for creating abuse cases utilizing CAPEC attack patterns. In their approach, words from use cases are mapped to technical terms (such as SQL, Schema, etc.), which are then used as keywords to search for CAPEC attack patterns. The method we propose is similar to the method proposed by Kaiya et al. in that both try to drive abuse cases based on use cases and CAPEC attack patterns. However, in our method, words from the use cases are not mapped to technical terms. Instead, they are directly used as keywords to search for relevant attack patterns from CAPEC. Kaiya et al. [9] did not discuss how words from use cases are mapped to technical terms. The relevance of the retrieved attack patterns to the use cases was also not discussed. In our method, a tool called TrAP was used to retrieve relevant attack patterns based on keyword search. Also, an abuse case report template was also provided to describe the generated abuse cases. Furthermore, we conducted a user study to evaluate the effectiveness of the proposed method, and gain insight on how a user (typically a non-expert in software security) select keywords and relevant attack patterns.

Yuan et al. [10] proposed an approach for non-experts to create meaningful abuse cases using threat modeling and CAPEC attack patterns [5]. Williams et al. [11] further evaluated this method using an online graduate Secure Software Engineering course. In [11], Microsoft Threat Modeling [12, 13] is first conducted to get keywords based on the threats generated and then the keywords were used to retrieve relevant attack patterns from CAPEC. The method we propose in this study is similar to that in [11], but does not require the threat modeling process which is considered as a design phase activity. The method that we propose can be applied to the requirements phase of the SDLC. Also, the user study described in this paper extends the user study in [11] in that an individual interview was conducted with the participants. The participants were asked questions such as how they selected the keywords from the use case description, and what they learned from applying the method.

3 Proposed Method for Deriving Abuse Cases

The proposed method for creating abuse cases based on use case descriptions and attack patterns is shown in Fig. 1 and the main steps described below:

1. *Select Keywords from Use Case Descriptions.* A use case description includes a summary of the use case, and a sequence of steps describing the interactions between a user and the system. We provided no guidelines or restrictions on how the participants should select the keywords from user case descriptions.

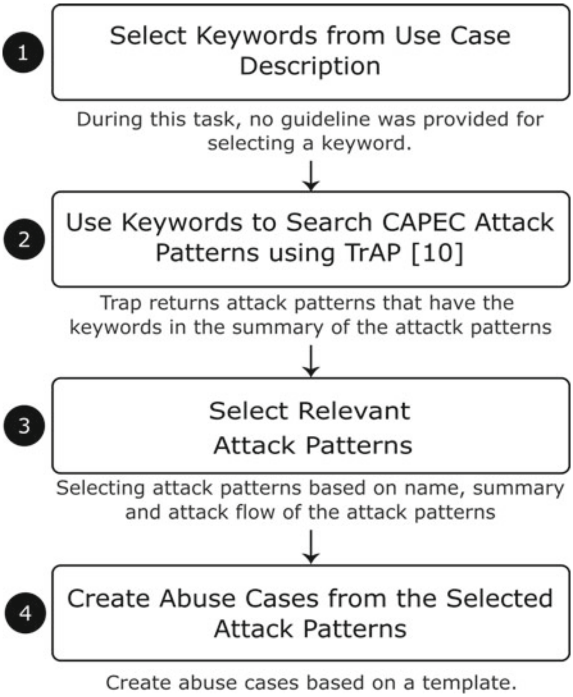

1
Select Keywords from Use Case Description
During this task, no guideline was provided for selecting a keyword.
2
Use Keywords to Search CAPEC Attack Patterns using TrAP [10]
Trap returns attack patterns that have the keywords in the summary of the attactk patterns
3
Select Relevant Attack Patterns
Selecting attack patterns based on name, summary and attack flow of the attack patterns
4
Create Abuse Cases from the Selected Attack Patterns
Create abuse cases based on a template.

Fig. 1 The method for creating abuse cases based on attack patterns

2. *Use Keywords to Search for CAPEC Attack Patterns*. In this step, the selected keywords were used to search for attack patterns from TrAP. TrAP [10] is a tool used to search for CAPEC attack patterns with a string search algorithm.
3. *Select Relevant Attack Patterns*. The third task is to select attack patterns that would affect the use case in respect to the software application. We suggested the participants use the title, description, and attack flow of an attack pattern where applicable to help determine and select a relevant attack pattern for the use case.
4. *Create Abuse Cases from the Selected Attack Patterns*. The final task is to create the abuse cases for the use cases using information from the relevant attack patterns. An abuse case template is provided for writing an abuse case. Some of the information comes from the attack pattern.

4 User Study of the Proposed Method

4.1 Participants and Materials

Twenty graduate students in an online graduate level Secure Software Engineering course at North Carolina A&T State University participated in this user study during the Spring 2017 semester. Prior to the study, the students learned the first seven chapters of the textbook [3] in this course. The materials provided to the participants of the user study were:

1. *Reading Materials*. Students were asked to read Chap. 8 in the textbook [3], which gives an overview of abuse cases and abuse case development.
2. *Assignment Instructions*. Instructions were given on how to use the proposed method to develop abuse cases. The instructions also include how to fill out the abuse case template and generate the assignment reports. How to use the TrAP tool is also included in the instruction.
3. *Software Requirement Specification (SRS)*. The participants were provided with a mock online shopping web application SRS document that had four use cases. These use cases were login, update account, checkout, and logout. These use cases were chosen because the participants are familiar with online shopping.

4.2 Procedure

As part of the course assessment, the participants were required to complete an abuse case development assignment. Prior to starting the assignment, the participants were provided with the reading materials, assignment instructions, and the SRS document. These materials were provided through Blackboard Learn.[1] The participants performed the following tasks:

1. Read Chap. 8 in the textbook [3] and the SRS document with the use case description.
2. Apply the proposed method to generate abuse cases for the use cases.
3. Complete the assignment report.
4. Voluntarily consent to participate in the online survey and individual interview that were approved by Institutional Review Board (IRB).

The participants were awarded 15 extra points for completing both the survey and the interview or an alternate class assignment was provided.

4.3 Data Collection

The following data collection methods were used:

1. *Assignment Report*. This includes the abuse cases that each participant created, the keywords used, and the attack patterns selected for deriving the abuse cases.

[1] Blackboard Learn, is a virtual learning environment and course management system developed by Blackboard Inc.

2. *Online Survey Questionnaire*. An online survey was designed and distributed to gain insights on the proposed method through participants' feedback, individual approaches the participants applied to complete the assignment, and obtain participants' background knowledge in software security.
3. *Individual Interview*. A follow up interview was conducted with each individual to understand how the participants selected the keywords, selected relevant attack patterns, and what knowledge they gained from applying the method.

4.4 Data Analysis

The participants' assignment reports, online survey responses, and interview responses were collected and then analyzed to evaluate the effectiveness of the proposed approach, and gain insight on how a participant selects keywords from use case descriptions, and select relevant attack patterns for developing abuse cases. Descriptive statistics methods were used to analyze the assignment reports. Qualitative research methods were used to analyze open-ended survey responses and the participants' interview responses.

5 Results and Discussion

Twenty (20) of the 21 participants completed the assignment. However, 5 of the 20 participants were excluded from the study because their assignment reports showed they did not follow the instruction to complete the assignment. From the 15 participants, eleven (11) completed the online survey and 7 completed the individual interview. Only one of the participants did not take any security related course(s) before taking this course. Eight (8) students mentioned that prior security courses they took helped them with the assignment. None of the participants had professional experience in software security.

5.1 Participants' Feedback on the Proposed Method

In the online survey, the participants were asked to rate the proposed method for creating abuse cases using a Likert scale of "Very Poor", "Poor", "Good", and "Very Good" along with an explanation of their rating. Six participants rated the method "Very Good", while four rated the method "Good", and one rated "Poor". Comments from the participants include:

1. The method is easy to follow and understand. It is a good way to create abuse cases.

2. Using TrAP to search CAPEC by keywords helps to find relevant attack patterns quickly.
3. The search results of TrAP are highly related to the use case. They can be used by developers to find attack patterns they are interested in.
4. CAPEC is a handy database to search for attack patterns.
5. This method increased student knowledge of security for building applications.

The participant who ranked this method poor mentioned that no examples were provided to help with filling out the fields of the abuse case template when the retrieved CAPEC attack pattern does not provide information for the fields.

The participants were asked to report the issues they encountered using the method. Some of the issues they reported include:

1. When the selected attack pattern did not have some of the information needed to fill out the abuse case template, the participants had to search for information outside CAPEC or use similar attack patterns to fill out the abuse case template.
2. Some use case descriptions do not have a lot of keywords.
3. Some keywords that are typically related to known risks did not return attack patterns.
4. A participant mentioned that he was uncertain about whether he followed the proposed method correctly. Another participant was unsure about creating the abuse cases in the beginning, but gained confidence after creating several abuse cases.

Some participants made the following suggestions to improve the method: (1) provide more detailed use case descriptions; and (2) provide detailed abuse case examples to demonstrate how to follow the proposed method.

The participants were also asked about the knowledge they gained through using this method. The participants reported that they increased knowledge on different security attacks and vulnerabilities, attack event flow, mitigation, risk analysis, and penetration testing.

5.2 Observation on Keyword and Attack Pattern Selection for Deriving Abuse Cases

Selecting Keywords from Use Cases' Description. The 7 participants who participated in the interview were asked how they selected the keywords from each use case description to search for attack patterns. Five participants selected words they thought were important to the use case, security related, or related to what the attackers would pay attention to. Two participants randomly selected noun/verbs as keywords.

Relevancy of the Derived Abuse Cases to the Use Cases. In the assignment instructions, we suggested that the "title", the "description", and the "attack event flow" sections of the attack pattern can be used to help determine whether an attack pattern

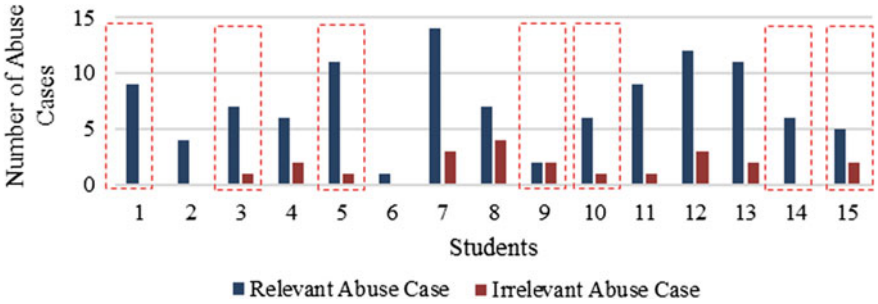


Fig. 2 The number of relevant and irrelevant abuse cases created by each participant

is relevant to the use case or not. The interview participants were asked to explain the approach they used to select attack patterns that are relevant to the use case. Participants selected attack patterns based on (1) the attack pattern description only; (2) if the attack pattern affects the use case/application; (3) the description and attack flow; and (4) just from the attack patterns that were returned. Figure 2 shows the number of relevant and irrelevant abuse cases generated by the 15 participants. The red highlighted rectangles are the results of the participants who completed the interview.

From Fig. 2, we can see that some interview participants did not include irrelevant abuse cases, while others included irrelevant abuse cases even though they mentioned that they used sections of the attack patterns to make their selection.

The attack patterns returned from TrAP is highly dependent on the keywords selected from the use case description. The use of security-related words to search for CAPEC attack patterns is more likely to return attack patterns that are relevant to the use case. Randomly using nouns and verbs that are not security-related may cause participants to select irrelevant attack patterns. Designing and implementing irrelevant security measures can take unnecessary time [14]. Therefore, security-related words should have higher priority as to being used as keywords for searching attack patterns.

5.3 Relationship Between Keywords and Attack Patterns

Keywords used to search for attack patterns. The collection of keywords used by the participants along with the top 5 keywords for each use case is listed below:

1. "*Authenticate User*": 12 different keywords were used for this use case. The top 5 keywords were: password, login, username, authenticate, and submit.
2. "*Check Out*": 10 different keywords were used for this use case. The top 5 keywords were: payment, purchase, transaction, shopping, and submit.
3. "*Update Customer's Account*": 7 different keywords were used for this use case. The top 5 keywords were: modify, account, Information, edit, and validate.

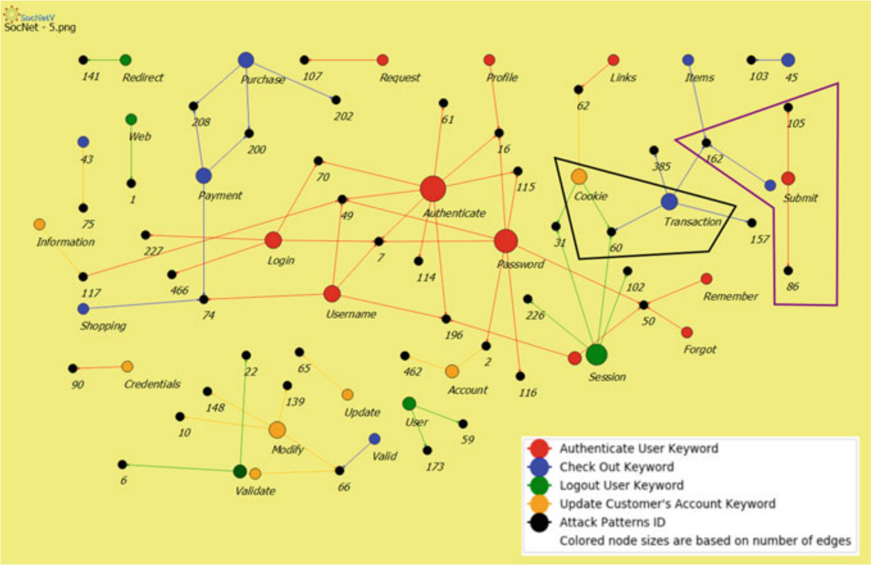


Fig. 3 Keywords versus attack patterns relationship graph

4. "*Logout User*": 7 different keywords were used for this use case. The top 5 keywords were: session, cookie, user, validate, and web.

A graph[2] was created to show the relationships between the keywords and the attack patterns retrieved using the keywords as shown in Fig. 3. In Fig. 3, the black nodes labeled with numbers represent attack pattern IDs, such as 66—SQL Injection; 49—Password Brute Forcing; 162—Manipulating Hidden Fields; 117—Interception. The nodes labeled with the keywords represent keywords for a use case which is color coded. Two noticeable relationships (enclosed in areas with purple boundary) can be observed in Fig. 3:

1. The keyword "*submit*" from the use case "Authenticate User" retrieved attack patterns 105 and 86. The same keyword "*submit*" for the use case "Check Out" retrieved attack pattern 162. This shows the same keywords can be used for different use cases, and they can retrieve different attack patterns for different use cases. See the nodes enclosed in purple Fig. 3.
2. The keywords "*cookie*" and "*transaction*" from the use case "Update Customer's Account" and "Check Out" retrieved attack pattern 60. This shows different keywords from the same use case can retrieve the same attack pattern for this use case. See the nodes enclosed in black Fig. 3.

From Fig. 3, we can also see: (1) The "Authenticate User" use case has more attack patterns compared to other use cases; and (2) the keywords are more related to

[2] http://socnetv.org/.

security. The interview responses show most of the participants were more familiar with the "Authenticate User" use case and used security related keywords from the use cases to search for attack patterns.

5.4 Limitation

This user study was limited by (1) the number of participants in the study was small for statistical analysis of data; (2) the assignment in the study did not specify the number of abuse cases the participants should derive; and (3) other course topics had to be taught before introducing the proposed method, which caused the study to be given close to the end of the semester. These limitations might have affected students' participation in the study and the statistical power of the data.

6 Conclusion

In this study, we proposed and evaluated a method for creating abuse cases based on use case description and CAPEC attack patterns. The results from the user study shows the proposed method is easy to follow, and helped the participants to gain more knowledge about software security. Also, based on how the participants selected keywords and the relevance of the selected attack patterns to the use cases, we suggest giving security related words priority in selecting keywords for search CAPEC attack patterns. We also observed the mapping relationships between keywords and the retrieved attack patterns. This mapping may imply possible inference to derive further relationships between keywords and attack patterns.

Our future work will include: (1) improving the proposed method based the participants' feedback; (2) create an ontology of keywords to augment the relationship between the use cases; (3) use other knowledge repositories such as Common Vulnerabilities and Exposures (CVE) and Common Weakness Enumeration (CWE) in the proposed method to build a recommender system that suggests relevant attack patterns and performs risk analysis of the attack patterns with respect to the use cases using Social Network Analysis centralities.

Acknowledgements This work is partially supported by National Science Foundation under the grant HRD-1332504 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Wei C. Sia: “misuse cases and abuse cases in eliciting security requirements”. System security: COMPSCI, vol 725
2. McDermott J, Fox C (1999) Using abuse case models for security requirements analysis. In: Proceedings of 15th annual computer security applications conference, 1999 (ACSAC’99), pp 55–64
3. McGraw G (2006) Software security: building security, vol 1. Addison-Wesley Professional
4. Hope P, McGraw G, Antón AI (2004) Misuse and abuse cases: getting past the positive. IEEE Secur Priv 2:90–92
5. CAPEC (2014) Classification (CAPEC)
6. Sindre G, Opdahl AL (2000) Eliciting security requirements by misuse cases. In: Proceedings 37th international conference on technology of object-oriented languages and systems, 2000 (TOOLS-Pacific 2000), pp 120–131
7. Alexander I (2002) Initial industrial experience of misuse cases in trade-off analysis. In: Proceedings of IEEE joint international conference on requirements engineering, 2002, pp 61–68
8. Pauli JJ, Engebretson PH (2008) Hierarchy-driven approach for attack patterns in software security education. In: Fifth international conference on information technology: new generations, 2008 (ITNG 2008), pp 1156–1157
9. Kaiya H, Kono S, Ogata S, Okubo T, Yoshioka N, Washizaki H et al (2014) Security requirements analysis using knowledge in capec. In: International conference on advanced information systems engineering, 2014, pp 343–348
10. Yuan X, Nuakoh EB, Beal JS, Yu H (2014) Retrieving relevant CAPEC attack patterns for secure software development. In: Proceedings of the 9th annual cyber and information security research conference. Oak Ridge, Tennessee, USA.
11. Yuan X, Nuakoh EB, Williams I, Yu H (2015) Developing abuse cases based on threat modeling and attack patterns. JSW 10:491–498
12. Microsoft threat modeling tool (2014) https://www.microsoft.com/en-us/download/details.aspx?id=42518. Accessed 28 Feb 2018
13. Owasp threat risk modeling. https://www.owasp.org/index.php/Threat_Risk_Modeling. Accessed 28 Feb 2018
14. Castañeda V, Ballejos L, Caliusco ML, Galli MR (2010) The use of ontologies in requirements engineering. Glob J Res Eng 10:2–8

An Efficient and Anonymous KP-ABE Scheme with Keyword Search

Tao Feng, Xiaoyu Yin and Chunyan Liu

Abstract In order to enhance the privacy preserving mechanism of cloud storage scheme and promote the cloud storage system in the enterprise, we propose a cloud storage scheme with multiple encryptions. The scheme adopts hybrid cloud structure and supports the direct revocation of users. The private cloud provided by the enterprise not only protects the attribute privacy in the access policy but achieves the pre-decryption mechanism reducing the decryption overhead for users. We introduce the identity authentication mechanism in the scheme and protect the users' identity though the homomorphic encryption over integers. Based on the Decisional Linear assumption and the integer approximate Greatest Common Divisor problem, the proposed scheme is secure in the random oracle model. The analysis shows that our scheme is more secure and practical.

Keywords Cloud storage · Privacy preserving · KP-ABE · PEKS

1 Introduction

1.1 Background and Related Works

The cloud storage system extending from the cloud computing has become a hotspot of internet applications. The cloud storage system not only saves the management costs of local data but also provides better redundancy backup and data recovery

T. Feng (✉) · X. Yin (✉)
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
e-mail: fengt@lut.cn

X. Yin
e-mail: 710547443@qq.com

C. Liu (✉)
School of Economics and Management, Lanzhou University of Technology, Lanzhou, China
e-mail: lcy_811@163.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_26

for users technically. However, the separation of data owner and data in the cloud storage service brings new security risks. There still are some problems such as privacy disclosure and over-calculation etc. in the current cloud storage schemes, which restricts the promotion and development of cloud storage in enterprise users [1]. The enterprise is an important kind of users in cloud storage applications. And the security requirement of data protection and privacy preserving mechanism is higher in the enterprise data sharing environment. For the convenience of enterprise users, the cloud storage system should be multifunctional and efficient.

Based on the identity-based encryption, Sahai and Waters proposed attribute-based encryption (ABE). With the development of ABE, the schemes that support fine-grained access control are classified into two categories: the key-policy attribute-based encryption (KP-ABE) and ciphertext-policy attribute-based encryption (CP-ABE). In view of the problem that ciphertext can't be searched in the attribute-based encryption cloud storage scheme, the researchers proposed cloud storage schemes with multiple encryptions. In literature [2], Xiong et al. proposed a secure and efficient scheme by combining the ciphertext-policy attribute-based encryption and homomorphic encryption. Both the outsourced data and the homomorphic key in the scheme need to be encrypted, and the two parts of the ciphertext are combined to generate the final ciphertext file. By combining symmetric searchable encryption (SSE) mechanism and CP-ABE, Wang proposed a cloud storage system with hybrid cloud architecture in literature [3]. The secondary encryption mechanism of the private cloud improves the security of the cloud service and realizes an efficient revocation of users. However, the access control mechanism in the scheme requires the user to submit the private key to the untrusted cloud server, which leads to the privacy leakage. In the key-policy attribute-based encryption research field, a searchable encryption scheme with access control reinforcement is proposed in literature [4]. The scheme adopts KP-ABE mechanism and implements the secondary filtering of search results. Users can filter out unwanted data from the search results. The filtering authority introduced in the scheme prevents the leakage of keywords. However, the schemes still has some problems such as privacy disclosure and large computation overhead. Han et al. [5] proposed an attribute-based encryption scheme that can transform the key-policy attribute-based encryption into attribute-based encryption with keyword search. In the scheme, users can make a more flexible search strategy though access structure, but the scheme lacking of the authentication of users' identity and revocation mechanism. Liu et al. further improved the verifiable attribute-based keyword search scheme proposed by Zheng et al. [6] and added the verification of search results to the cloud storage system by introducing signature mechanism [7]. But the attribute set is easily steal by untrusted cloud servers in this scheme. Zhu introduced the KP-ABE into the public key encryption with equality test (PKEwET) scheme and proposed a new cloud storage scheme [8]. The scheme supports the fine-grained authorization and can detect whether the ciphertext encrypted by different public keys contains same plaintext information. Whereas this scheme still lacks of revocation mechanism. In current research of key-policy attribute-based encryption, the privacy preserving and revocation mechanism need to be improved further.

1.2 Our Contribution

We combines the KP-ABE, the public key encryption with keyword search (PEKS) and homomorchic encryption over integers to implement a practical cloud storage scheme for enterprise users. Firstly, we adopt the hybrid cloud structure to improve the privacy preserving mechanism of attribute set. Secondly, the homomorchic encryption over integers is introduced to solve the problem of users' identity disclosure. Finally, we add the direct revocation to the hybrid cloud storage scheme and achieve the pre-decryption of users with the assist of private cloud.

2 System Model

We propose an efficient and anonymous KP-ABE scheme with keyword search. In this section, we will describe the basic structure as well as threats model of the proposed scheme.

2.1 Threats Model

In the proposed scheme, the private cloud provided by the enterprises and the key generator are fully trusted. The public cloud is honest but curious. It will abide by the protocol and return the searched ciphertext to users. However it may steal the privacy information of system users. Users are semi-trusted and they may collude to get the data without permission.

2.2 Basic Structure

There are five participating entities in our scheme including data owner, users, public cloud, private cloud, key generator. The basic structure of scheme is shown as Fig. 1.
Data owner: The data owner encrypts the keyword and data. The ciphertext related to the attribute is send to private cloud to achieve access control of users. The ciphertext of keyword and outsourcing data are submitted to the public cloud.
Users: Users need to register themselves on the private cloud. They get the private key from the key generator by submitting the access structure. Before requesting the ciphertext, users encrypt the private key and then send it to the private cloud for pre-decryption. The trapdoor generated with the user key and the ciphertext of their identity are sent to the public cloud to get the shared data. After receiving the ciphertext, users recover the plaintext of shared data with the user key.

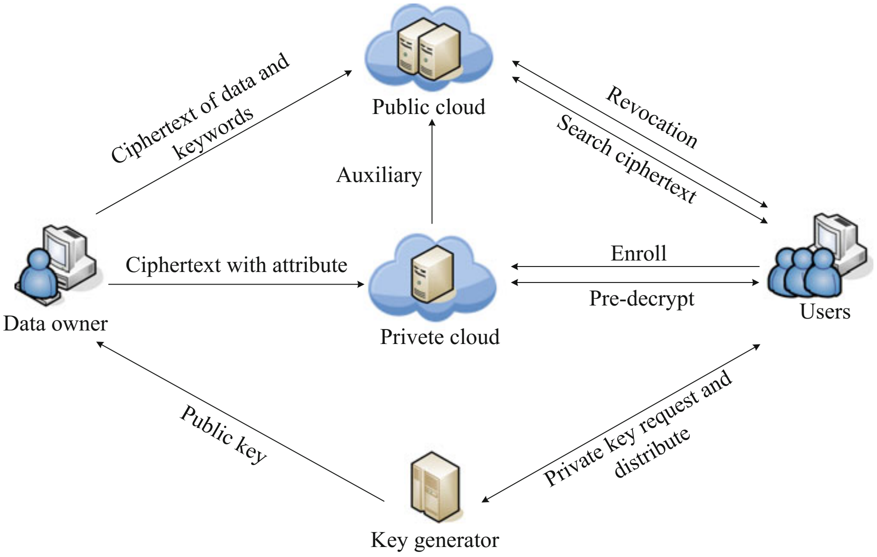


Fig. 1 Efficient and anonymous KP-ABE scheme with key word search

Public cloud: The public cloud stores the ciphertext and responds to users' search requests. If user's identity is verified, the public cloud will perform the search operation. And then it returns the pre-decryption key and the searched ciphertext to users. In the phase of revocation, it verifies the user's identity and deletes the corresponding item from user list.

Private cloud: The private cloud responds to users' enroll request by generating the identity token of user and the corresponding user key. Besides, it provides the pre-decryption computation for users and encrypts the identity token with homomorchic algorithm. The pre-decryption key is transmitted to the public cloud along with the ciphertext of user's identity.

Key generator: The initialization of system is executed by key generator, which generates the public parameters and the master key. The key generator also calculates the corresponding private key after receiving user's access structure.

3 Concrete Algorithm

In this section, we will give the algorithm structure of the scheme, which is composed by ten algorithms.

Setup: The key generator chooses a bilinear map $e : G \times G \rightarrow G_T$, where G and G_T are cyclic prime order p. The generator of G is g. Let $H_1 : \{0, 1\}^* \rightarrow G$ and $H_2 : \{0, 1\}^* \rightarrow Z_p$ be the hash function and pick up random number $a, b, c \in Z_p$.

We define that public key of system is $PK = \{H_1, H_2, g, g^a, g^b, g^c, e(g, g)^{ac}\}$ and the master key is $MK = \{a, b, c\}$.
Enroll: The private cloud picks the user key (*UK*) for registered user and generates the corresponding identity token (*GID*). Select two secure large prime numbers *P* and *Q*. Then, calculate $N = P \times Q$ and use the p as key. The ciphertext of user's identity is given as $C_U = (M_{GID} + P \times a_1) \mod N$ where a_1 is a random number.
Encrypt: The algorithm picks up random numbers $r_1, r_2 \in Z_p$ and the ciphertext related to the attribute is $C_{att} = \{Atts, C_0 = g^{r_2}, C_i = H_1(atti)^{r_2}\}$ where *Atts* is the attribute set. The ciphertext of keyword *W* and data is given as:

$$C = \{C_M = Me(g, g)^{ac \cdot r_2}, C_1 = g^{c \cdot r_1}, C_2 = g^{b \cdot H_2(W) \cdot r_1} g^{a \cdot (r_1 + r_2)}\}$$

Before outsourcing the decryption, the private key needs to be encrypted with the user key as $C_{SK} = (SK_1^{UK}, SK_2^{UK}) = (g^{q_j(0) \cdot UK} H_1(attj)^{t \cdot UK}, g^{t \cdot UK})$. And the identity of user is submitted to public cloud is encrypted as $C_{GID} = (m_{GID} + P \times a_2) \mod N$ where a_2 is a random number.
KeyGen: The algorithm chooses a polynomial $q_j(0)$ for each node j in access tree T_p and sets $q_{j(0)} = p_{parrent(j)}(index(j))$. For the root node r, let $q_r(0) = ac$. Picks up a random number *t* and calculates $SK_1 = g^{q_j(0)} H_1(attj)^t$, $SK_2 = g^t$. The private key is given as $SK = (SK_1, SK_2)$.
Trapdoor: In this algorithm, the trapdoor is generated with the user key (*UK*) and the calculation is $TK_1 = g^{a \cdot UK} g^{b \cdot H_2(w) \cdot UK}$, $TK_2 = g^{c \cdot UK}$, $TK = (TK_1, TK_2)$.
Pre-decrypt: To access control the test algorithm is:

$$E_j = \frac{e(C_{SK_1}, C_0)}{e(C_{SK_2} C_i)} = \frac{e(g^{q_j(0) \cdot UK} H_1(attj)^{t \cdot UK}, g^{r_2})}{e(g^{t \cdot UK}, H_1(atti)^{r_2})} = e(g, \mathrm{g})^{q_j(0) \cdot UK \cdot r_2}$$

The pre-decryption key is $E_{root} = e(g, g)^{ac \cdot UK \cdot r_2}$.
Verify: After receiving the ciphertext of user's identity, the verifying is executes as follows where $N = P \times Q$ and a_t is the random number.

$$\begin{aligned} VE &= ((C_U - C_{GID}) \times Q \times a_t) \mod N \\ &= ((M_{GID} - m_{GID}) \times Q \times a_t + (a_1 - a_2) \times P \times Q \times a_t) \mod N \end{aligned}$$

If $M_{GID} = m_{GID}$, $VE = 0$ and user's identity is authenticated. If $VE > 0$, the public cloud aborts and returns a message that the search of ciphertext is stopped.
Test: The search algorithm tests the equation $e(C_1, TK_1) \cdot E_{root} = e(C_2, TK_2)$. If the searched key word w is same as the key word of outsourcing data W, the equation will be established.

The correctness of equation:

$$e(C_1, TK_1) \cdot E_{root} = e(g^{cr_1}, g^{a \cdot UK} g^{b \cdot H_2(w) \cdot UK}) \cdot e(g, g)^{ac \cdot UK \cdot r_2} = e(g, g)^{acUK(r_1 + r_2) \cdot} \cdot e(g, g)^{bcr_1 H_2(w) \cdot UK}$$

$$e(C_2, TK_2) = e(g^{b \cdot H_2(W) \cdot r_1} g^{a \cdot (r_1 + r_2)}, g^{c \cdot UK}) = e(g, g)^{ac \cdot UK \cdot (r_1 + r_2)} \cdot e(g, g)^{bcr_1 \cdot H_2(W) \cdot UK}$$

Decrypt: the algorithm recovers the searched data with the pre-decryption key and user key as:

$$M = \frac{C_M}{E_{\mathrm{root}}^{1/UK}}$$

Revoke: In the phase of revocation, users submit the ciphertext of their identity and then the public cloud executes verify algorithm. If $VE = 0$, the public cloud removes the corresponding item from the user list.

4 Security and Performance Analysis

4.1 Security Proof

Our scheme is based on the KP-ABE scheme proposed in literature [6]. Given the Decisional Linear (DL) assumption, the scheme is selectively secure against chosen-keyword attack in the random oracle model. The appendix A in [6] has given the theorem and complete proof of security. In our extended algorithm, the user key generated randomly and the random number doesn't reduce the security of scheme in the process of improving the privacy preserving mechanism. Besides, the security of homomorphic encryption over integers algorithm [9] used in this paper is based on the integer approximate Greatest Common Divisor intractable problem and the safety is satisfactory.

4.2 Privacy Preserving Analysis

Data owner. The outsourcing data is encrypted with KP-ABE, the public key cryptosystem which has higher security. In order to further protect the data in cloud, the identity authentication mechanism is introduced during the search process to prevent malicious users from illegally stealing data. In addition, the access control of the scheme is completed by the cooperation of public cloud and private cloud. Besides, the attribute set specified by the data owner doesn't need to be uploaded to the untrusted public cloud, thereby avoiding the privacy leakage problem of attribute set.

User. The scheme adopts the homomorphic encryption technology to protect the privacy of the user's identity. In the process of data requiring and revocation, user's identity is exist in ciphertext form. In the pre-decryption phase, the user first encrypts the private key using the user key (UK) and then submits the ciphertex of private key for calculation, which protects users' privacy effectively.

Table 1 Comparison

Scheme	Hybrid cloud	Pre-decrypt	Keyword search	Revocation	Attribute privacy	Identity privacy
[2]	×	√	√	×	×	×
[3]	√	×	√	√	×	×
[4]	×	×	√	×	×	×
[5]	×	×	√	×	×	×
[6]	×	√	√	×	×	×
our	√	√	√	√	√	√

4.3 Performance Analysis

In this paper, the full homomorphic encryption over integers mechanism used to encrypt user's identity has the simpler key and less time complexity. What's more, the pre-decryption mechanism is adopted to reduce the computation overhead of decryption for users. Because the complex bilinear pairing operations are outsourced to the private cloud, users can recover the ciphertext by a simple calculation on the computing constrained device such as handheld devices.

5 Comparison

We compare the proposed scheme with some other cloud storage schemes as shown in Table 1.

By combing the ciphertext-policy attribute-based encryption with the homomorphic encryption and PEKS respectively, the scheme in literature [2, 3] achieves ciphertext search operation. In the schemes adopting KP-ABE and searchable encryption such as [4–6], there are still some problems such as lacking of revocation mechanism and inadequate privacy preserving mechanism. The scheme in this paper uses hybrid-cloud mechanism to protect the attribute in access structure and identities of users. The scheme supports direct revocation of user and decreases the cost of decryption for users.

6 Conclusion

Due to the complexity of network environment, the privacy disclosure problem in cloud storage services has drawn widespread attention of researchers. This paper proposes an efficient and anonymous KP-ABE scheme to generalize cloud storage system to enterprise users. The multiple encryption mechanisms are introduced to perfect the privacy preserving mechanism and optimize the performance of system. The hybrid-cloud structure is introduced into the searchable KP-ABE scheme, which realizes the pre-decryption mechanism. Besides, the proposed scheme achieves direct revocation of users and adopts identity authentication to protect the outsourced data further. Analysis shows that our scheme is more practical and more secure.

Acknowledgements This study was supported by the National Natural Science Foundation of China (No. 61462060, No. 61562059), Regional Science Foundation Project (No. 61762060).

References

1. Feng T, Yin XY (2016) Research on privacy preserving mechanism of attribute-based encryption cloud storage. Chin J Netw Inf Secur 2(7):8–17
2. Xiong AP, Gan QX, He XX et al (2013) A searchable encryption of CP-ABE scheme in cloud storage. In: International computer conference on wavelet active media technology and information processing, pp 345–349
3. Wang Q, Zhu Y, Luo X (2015) Multi-user searchable encryption with coarser-grained access control without key sharing. In: International conference on cloud computing and big data. IEEE, pp 119–125
4. Kaci A, Bouabana-Tebibel T (2015) Access control reinforcement over searchable encryption. In: IEEE international conference on information reuse and integration. IEEE, pp 130–137
5. Han F, Qin J, Zhao H et al (2012) A general transformation from KP-ABE to searchable encryption. In: Cyberspace safety and security. Springer, Berlin, Heidelberg, pp 107–115
6. Zheng Q, Xu S, Ateniese G (2015) VABKS: verifiable attribute-based keyword search over outsourced encrypted data. IEEE INFOCOM. IEEE, pp 522–530
7. Liu P, Wang J, Ma H et al (2015) Efficient verifiable public key encryption with keyword search based on KP-ABE. In: Ninth international conference on broadband and wireless computing, communication and applications. IEEE, pp 584–589
8. Zhu H, Wang L, Ahmad H et al (2017) Key-policy attribute-based encryption with equality test in cloud computing. IEEE Access 99:1–1
9. Meiyun LI, Jian LI, Huang Chao (2012) A credible cloud storage platform based on homomorphic encryption. Netinfo Secur 12(9):35–40

A DNS RPZ Firewall and Current American DNS Practice

Norman Wilde, Lauren Jones, Robert Lopez and Travis Vaughn

Abstract Many varieties of internet attack make use of the Domain Name System (DNS) at some point. Response Policy Zones (RPZ) is a reputation-based DNS firewall technology intended to obstruct the use of the DNS by malicious actors. We report on a study that compared the blocking behavior of a freely available RPZ-enabled DNS service with the blocking behavior of other DNS services available in the United States. We were surprised to find that only the RPZ-enabled server was doing any significant blocking of malicious domains. Since our study was carried out, this free RPZ-enabled service has been made more widely available as *Quad9* (https://quad9.net/).

1 Introduction

Many forms of internet attack make use at some point of the Domain Name System (DNS). Often attacks begin by luring a user to a website that distributes malware. Attackers create and register these sites by the thousands every day [1]. Malware domain names are selected to seem innocuous or familiar so that even a cautious user may be deceived by a phishing email or an innocently appearing web link. In other attacks the malware in already-compromised computers uses DNS to “call home” to domains that map to a botnet controller. For example, the Conficker worm attempted to connect at random to such domains to find a server from which it could download updates to itself [2, 3].

N. Wilde (✉) · L. Jones · R. Lopez · T. Vaughn
University of West Florida, Pensacola, FL, USA
e-mail: nwilde@uwf.edu

L. Jones
e-mail: lnj6@students.uwf.edu

R. Lopez
e-mail: rl34@students.uwf.edu

T. Vaughn
e-mail: tlv18@students.uwf.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_27

One way to improve internet security could be to obstruct the use of DNS by attackers. Response Policy Zones (RPZ) is a technology that has been proposed to block access to known malware sites from end-user computers. Several implementations of RPZ exist, but the technology is still not widely known or used.

In this paper we describe an experimental study carried out in April of 2017 to compare a free RPZ-enabled DNS service with other DNS services that are typical of those available in the United States. The purpose of the study was to compare the security protections offered by the RPZ implementation with those offered in current American practice.

In the next two sections we discuss DNS RPZ and related work on DNS firewalls. Sections 4 and 5 provide an overview of the experimental study and its results. (A more detailed description is available as a technical report of the Security and Software Engineering Research Center [4]). Section 6 summarizes our conclusions.

2 DNS and RPZ

The DNS is a hierarchical distributed database that allows mapping between domain names (e.g. www.example.com), and the numeric IP address of the computer that hosts that domain (e.g. 172.16.254.1). Normally, every internet-connected computer is provided with the IP address of at least one DNS server. Queries sent to that server may either be answered directly from cached information, or else the query is forwarded to root servers for a hierarchical recursive search until an authoritative response is found [5].

RPZ enables the creation of a rapidly updated “DNS firewall” at any DNS server. The server accesses one or more reputation feeds from security service providers that provide up-to-date data on known malicious domain names. Information from the reputation feeds is overlaid on information from the global DNS allowing the server to provide alternative responses to queries [6].

For example, if an end user’s computer requests the IP address for a domain name that has been tagged as malicious by a reputation feed, the DNS server may return NXDOMAIN (no such domain) instead of an IP address [7]. Without the IP address the user’s computer cannot access the malicious site and the attack is blocked, at least temporarily.

3 Related Work

RPZ is a particular case of a DNS-focused reputation-based approach to internet security. Such approaches involve creating reputation data for domain names or IP addresses and then using that data in an intrusion detection system or a firewall [2].

Most of the academic research on DNS firewall approaches has addressed the problem of creating the necessary reputation data. A main difficulty lies in keeping lists current in view of the very large number of new registrations encountered every day [1].

Rahbarinia, Perdisci, and Antonakakis proposed a detection system called Segugio for use within a large ISP network. It collects and analyzes information on the DNS network traffic to identify domains that are likely to be involved in malware control [8].

Hao, Kantchelian, Miller, Paxson, and Feamster reported on a proactive reputation system called PREDITOR that analyzes characteristics of a domain name that can be assessed at the time of registration to identify new malicious registrations quickly [1].

We are aware of only one paper that studied a working RPZ implementation. Connery described a trial at the Technical University of Denmark. Over the 4 week period of the trial about 5000 attempted contacts to dangerous domains were prevented. Participants reported no loss of productivity or inappropriate filtering [7].

Standardization of RPZ apparently is still an ongoing process. The first formal description of RPZ seems to have been the 2010 ISC Technical Note by Vixie and Schryver [9]. There is an Internet Engineering Task Force draft dated October 11, 2016 but it is still classified as a "work in progress" [10]. There is also a reference implementation of RPZ in the BIND 9.8 DNS server and there are a number of reputation feeds and internet firewall services available [6].

4 Experimental Study

In the experiment we ran a large number of queries in 7 different DNS environments, each being a combination of a client that made the queries and a specific DNS server that received and responded to the queries, possibly with the aid of other nameservers with which it could communicate.

The 7 DNS environments used were:

- An RPZ environment with a forwarding server deployed on an instance (virtual machine) in the Amazon Web Services EC2 cloud. This server forwarded all queries to a free RPZ-enabled DNS platform offered by the Global Cyber Alliance (GCA) in partnership with the Packet Clearing House (PCH) [11]. This GCA environment drew on 10 reputation feed sources. The client ran on a different instance in the same EC2 region.
- Three DNS service environments widely available in the United States and elsewhere. These used well known DNS servers provided by Google, Dyn, and OpenDNS respectively. These servers have been classed among the top public DNS services [12]. The client ran on the same EC2 instance used with the RPZ environment.
- Three local DNS environments available in the state of Florida, USA. One was a DNS environment at the University of West Florida (UWF). It may be taken

as a representative of organizations that have their own DNS infrastructure. The client for this environment was a faculty office computer. Two other services were the DNS environments provided by Internet Service Providers (ISPs) AT&T and Cox Communications. The clients for these environments ran on home computers owned by study participants and sent queries direct to the ISP's DNS server.

Each environment was queried using nine lists of domain names obtained from various sources. These lists fell into three categories:

- Two blacklists from reputation services that can be expected to consist mainly of malicious sites. One list came from the Adblock Plus web site [13] and one from the DNS-BH (Black Hole DNS Sinkhole) web site [14]. There was quite a lot of overlap between these two lists.
- Three lists "scraped" by web crawling links from sites dealing in pornography or free video downloads; such sites are reputed to have links to dangerous pages. As starting points for scraping we used https://www.pornhub.com/, https://www.heavy-r.com/ and https://movies4star.com/.
- Four lists which may be more typical of normal internet usage. Three of these lists were scrapings from well-known portals: https://www.yahoo.com/, http://www.cnn.com/ and https://www.bankofamerica.com/. We also obtained an anonymized log of two days of internet use by employees at a local educational organization.

5 Experimental Results

The query runs were made over a period of 6 days, April 14–19, 2017 and gave a total of 42,033 query responses. We analyzed these responses to look for evidence of blocking of malicious domains.

The most likely way for a DNS firewall to protect an end user is by returning the NXDOMAIN response to a query. Unfortunately, this response may also mean that the domain simply does not exist. To distinguish between these two cases, we sent the same queries to a baseline environment. This was a "plain vanilla" BIND server known not to be doing any blocking, so when it responded NXDOMAIN we could be sure that the domain really did not exist, at least at the moment when the baseline queries were run. We removed all queries to domain names that baseline reported as NXDOMAIN which left us with a total of 37,975 query responses. Remaining NXDOMAIN query responses are either:

- Security blocking of potentially malicious domains, or
- Random error caused by a domain name vanishing between the time of the baseline run and the environment's run.

The second category should be small so any substantial number of NXDOMAIN responses from any DNS environment would indicate filtering for security reasons.

The results are shown in Table 1. We can easily see that only the RPZ GCA service is doing any significant blocking by returning NXDOMAIN for malicious

Table 1 NXDOMAIN query responses

Sample list	#	DNS environments						
		RPZ	University	DNS Services			ISPs	
		GCA	UWF	Google	Dyn	OpenDNS	CoxISP	ATT
AdBlock	3533	2948	15	11	7	12	18	10
DNSBlackHole	3575	2996	15	19	2	19	19	4
pornhub	4811	1	1	1	0	2	1	0
heavyr	2329	2	0	0	0	0	0	0
movies4 stars	4676	3	1	0	0	1	1	0
yahoo	4896	9	0	0	0	0	0	0
cnn	4928	12	0	0	0	0	1	0
Bank of America	4886	16	0	0	0	1	0	0
InternetLog	4341	7	3	3	0	0	4	4
Total	37,975	5994	35	34	9	35	44	18

sites. RPZ GCA returned NXDOMAIN 5994 times and these returns are concentrated in the reputation blacklist domain name lists (AdBlock, DNSBlackHole). No other environment returned more than 44 NXDOMAINs.

Instead of returning NXDOMAIN, another possible blocking strategy is to return the IP address of a safe “walled garden” site to redirect users to a warning message or some other safe content. Identifying walled garden blocking is a bit difficult since the DNS service may not be returning the same IP address all the time; it is likely that several addresses would be used thus balancing the redirected load over several physical servers.

To look for walled garden blocking, we took the set of domain names that had been blocked by the RPZ GCA service and thus have a high risk of being malicious. We hypothesized that any DNS service using a walled garden would:

- Redirect to the garden at least 25 of these high-risk domain names (about 0.5% of the total set).
- Use a garden IP address range having the same first three octets, that is, the same first 24 bits.

Thus, for each DNS service we examined all the 24-bit IP ranges that showed up in more than 25 query responses to the high-risk list. However, none of these responses showed any sign of being anything other than real responses from the authoritative name server.

We conclude that the RPZ GCA service is blocking large numbers of potentially malicious sites, but the other services are not doing any blocking, neither by returning NXDOMAIN nor by walled garden redirection.

6 Conclusions

In this study we have contrasted the blocking behavior of DNS services typical of those in use in the United States with that of an RPZ enabled DNS service. We were rather surprised to find that none of the other DNS services seemed to be making any attempt to block malicious web sites. Since our experiments in April of 2017, the RPZ-enabled DNS service has been made freely available as a public service named *Quad9* [15].

RPZ complements, rather than replaces, other reputation-based filtering strategies currently used to enhance internet safety. Web browsers typically provide a warning if the user attempts to navigate to a website that is on a black list (e.g. for Firefox see [16]). Google Safe Browsing is a free service that enables such warnings in email and browser clients by letting them check URLs against lists of unsafe web sites [17].

However, attacks sometimes manage to bypass such browser-based defenses and may also penetrate via channels other than browsers. RPZ can provide an additional layer of defense and may be especially useful in obstructing "call home" behavior after an attack gains a lodgment. Malware that cannot connect to its command server may have difficulty updating itself or exfiltrating data.

DNS firewalls using RPZ would seem to have considerable potential to make the internet somewhat safer. Organizations that have a large number of computer users could well consider adding RPZ to their defenses. It should at least make the task of the attacker more difficult and perhaps go some way to mitigate the current imbalance between cyber-attack and cyber-defense.

Acknowledgements Data collection for the study described in this report was carried out in a course project by undergraduate students in computing at the University of West Florida. We would like to thank all class members for their participation. We would also like to thank Adnan Baykal of the Global Cyber Alliance for his guidance.

References

1. Hao S, Kantchelian A, Miller B, Paxson V, Feamster N (2016) PREDATOR: proactive recognition and elimination of domain abuse at time-of-registration. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (CCS'16). ACM, New York, NY, USA, pp 1568–1579. https://doi.org/10.1145/2976749.2978317
2. Antonakakis M, Perdisci R, Dagon D, Lee W, Feamster N (2010) Building a dynamic reputation system for DNS. In: Proceedings of the 19th USENIX conference on security (USENIX Security'10). USENIX Association, Berkeley, CA, USA
3. Shin S, Gu G (2010) Conficker and beyond: a large-scale empirical study. In: Proceedings of the 26th annual computer security applications conference (ACSAC'10). ACM, New York, NY, USA, pp 151–160. https://doi.org/10.1145/1920261.1920285
4. Jones L, Lopez R, Vaughn T, Wilde N (2017) Measuring the impact of DNS resource policy zones, Security and Software Engineering Research Center Technical Report S2ERC-TR-327, http://www.SERC.net, http://www.normanwilde.net/publications/TecRpt327/index.html

5. Liu C, Albitz P (2009) DNS and BIND, 5th edn. O'Reilly. Ebook ISBN 978-0-596-10572-3
6. DNS response policy zones. https://dnsrpz.info/
7. Connery HM (2013) Response policy zones: history, overview, usage, and research. https://dnsrpz.info/RPZ-History-Usage-Research.pdf
8. Rahbarinia B, Perdisci R, Antonakakis M (2016) Efficient and accurate behavior-based tracking of malware-control domains in large ISP networks. ACM Trans Priv Secur 19(2). https://doi.org/10.1145/2960409
9. Vixie P, Schryver V (2010) DNS response policy zones (DNS RPZ), ISC-TN-2010-1-B3, Dec 2010 (Draft 3). https://ftp.isc.org/isc/dnsrpz/isc-tn-2010-1.txt
10. Vixie P, Schryver V (2016) Response policy zones: draft-vixie-dns-rpz-00. https://tools.ietf.org/html/draft-vixie-dns-rpz-00
11. Global Cyber Alliance. Internet immunity: protecting users and networks via DNS. https://www.globalcyberalliance.org/wp-content/uploads/GCA-DNS-infrastructure.pdf
12. Meisel M (2011) Top public DNS resolvers compared. http://www.circleid.com/posts/20110407_top_public_dns_resolvers_compared
13. Adblock plus. https://adblockplus.org/
14. DNS-BH—malware domain blocklist. http://www.malwaredomains.com/
15. Quad9—internet security & privacy in a few easy steps. https://quad9.net/
16. How does built-in phishing and malware protection work. https://support.mozilla.org/en-US/kb/how-does-phishing-and-malware-protection-work
17. Google safe browsing. https://safebrowsing.google.com/

SNMP-Based Detection of VLAN Hopping Attack Risk

Kwangjun Kim and Manhee Lee

Abstract Virtual local area network (VLAN) is commonly used to divide a big network into several small network segments. Also, many adopt VLAN for dissecting LANs in order to prevent communications between different VLANs for security and management purposes. It is known that inserting an additional VLAN tag into Ethernet frames, referred to as VLAN hopping attack, can bypass the VLAN-based network separation. There are two preconditions for the attack. The first condition is that a hacker needs to know the destination's VLAN identification number and the second condition is that the attacking system needs to be connected a switch's trunk port that is used for connecting a switch. In this study, we propose an SNMP (Simple Network Management Protocol)-based detection method to effectively find a port and an MAC address that meet the second condition before a VLAN hopping attack begins. Since SNMP is implemented by most network components, our method can be easily deployed to the current VLAN networks.

Keywords Virtual LAN · VLAN hopping attack · SNMP

1 Introduction

If a company wants to divide a large network into smaller networks for security and management purposes, the virtual LAN (VLAN) technology can be used to efficiently partition the network [1, 2]. It creates multiple broadcast domains by dividing a single 2-layer network. This technology can be used as an economical network partitioning technique because it can separate the network by only changing the configuration of the existing switch without purchasing additional equipment.

K. Kim · M. Lee (✉)
Department of Computer Engineering, Hannam University, Daeduck-gu, Daejeon, Korea
e-mail: manheelee@hnu.kr

K. Kim
e-mail: kimkwangjun.kr@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_28

Although VLANs prevent packets from reaching different VLANs, international standard IEEE 802.1Q VLANs are known to allow network packets to arrive between different VLANs [3]. The root cause of this vulnerability is that the Ethernet packets that the PC sends to the switch should not contain the VLAN tag, so the normal access port of the switch drops these packets, but the trunk port successfully forwards them since the trunk port is designed to do so. Hackers devised an attack that bypasses VLAN-based network separation by intentionally inserting a VLAN tag to Ethernet frames. This attack is very important for security. This is because the ability to reach packets between different VLAN segments means that data can be leaked or attacks from the Internet are possible.

A simple countermeasure is not to connect a normal PC to the trunk port, but this is not a complete solution. This is because Cisco provides the Dynamic Trunking Protocol (DTP), which allows a connected system to dynamically change the access port to a trunk port [4, 5]. For better security, it is advised to configure the access port not to be changed to the trunk mode [6], but most of the switch ports are vulnerable to attack because the DTP feature is enabled by default [7]. To our best knowledge, there is no effective solution to detect or prevent it.

In this paper, we propose a method to utilize SNMP provided by most switches [8]. Our main idea is to periodically monitor the status of the entire switch port, i.e., whether it is a trunk or an access port, and then check that the PC is connected to that port. Since the situation where PC is connected to the trunk port is an important precondition for VLAN hopping attack, if we detect this situation quickly and immediately send an alert to the administrator, we can reduce the time for the VLAN hopping attack to occur. This paper shows that this method is possible with SNMP.

This paper is composed as follows. In Sects. 2and 3, we explain the threat model and how VLAN hopping attacks occur. In Sect. 4, we present an SNMP-based method to detect the situation where VLAN hopping attack can occur. We conclude with Sect. 5.

2 Threat Model

First, we briefly describe the threat model that we will discuss in this paper. The PC connected to the switch and switch itself is physically secure. The switch and the victim PC are not compromised by hackers. A hacker PC was compromised by a remotely connected hacker. The current network is separated by VLAN and is operating normally. The goal of the hacker pc is to send packets reaching the victim PCs belonging to different VLANs. Our goal in this threat model is to quickly detect this situation and help prevent future attacks.

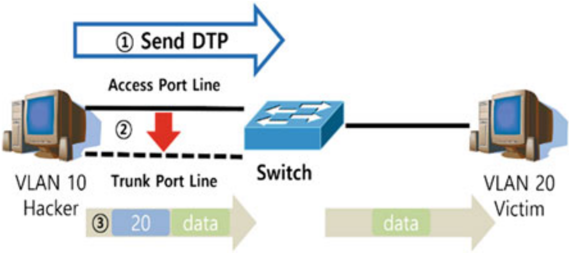


Fig. 1 VLAN hopping attack after changing to trunk port

3 VLAN Hopping Attack

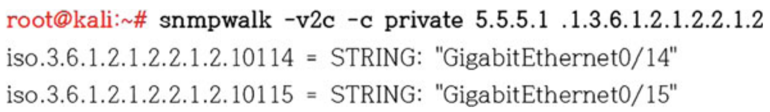
VLAN hopping attacks can occur in the following stages: The first step is to change the mode of the switch connected to the hacker PC from the access port to the trunk port mode. This is because if the access port is set, Ethernet packets containing the VLAN tag are dropped from the access port due to the wrong packet format. As shown in [3, 6], the attacker sends DTP packets that can change the port of the switch from access mode to trunk mode. When DTP packets are received, the switch whose DTP function is enabled changes the state of the port where the DTP packet arrives from Access mode to Trunk mode. The attacker then inserts a tag with the VLAN ID to which the packet is sent into the normal packet, and the packet is forwarded to that VLAN. In Fig. 1, a VLAN ID of 20 is inserted into the packet and the packet can be transmitted to the victim PC belonging to VLAN 20.

Since DTP is a proprietary protocol from Cisco, other companies' switches could seem safe. However, when a PC is connected to the trunk port by mistake of the network administrator, this problem can happen.

4 SNMP-Based Detection of VLAN Hopping Attack Risk

We propose a method to detect this attack at an early stage using SNMP, a standard protocol for managing network components. The management information base (MIB) is a database used to remotely manage nodes specified by the agent using SNMP. This database is a hierarchical tree structure, with each entry specified via an Object Identifier (OID) [9]. OID, also called object identifier, is formally defined and standardized by the International Telecommunication Union and ISO, IEC.

We propose to detect this attack situation with the following steps. First, we read *ifDescr* (OID: 1.3.6.1.2.1.2.2.1.2) from each entry, *ifEntry*, in the *ifTable* of the SNMP MIB-2 Interfaces to obtain information about each interface. It is used to check port numbers of the interfaces to be searched [10]. Figure 2 shows an example: the type of two ports is Gigabit Ethernet and port numbers are 14 and 15. Its network testbed environment is depicted in Fig. 3.



```
root@kali:~# snmpwalk -v2c -c private 5.5.5.1 .1.3.6.1.2.1.2.2.1.2
iso.3.6.1.2.1.2.2.1.2.10114 = STRING: "GigabitEthernet0/14"
iso.3.6.1.2.1.2.2.1.2.10115 = STRING: "GigabitEthernet0/15"
```

Fig. 2 Interface information from *ifEntry*

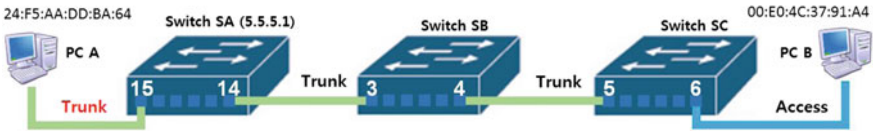
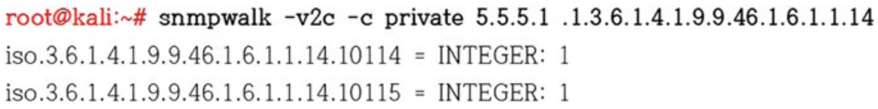
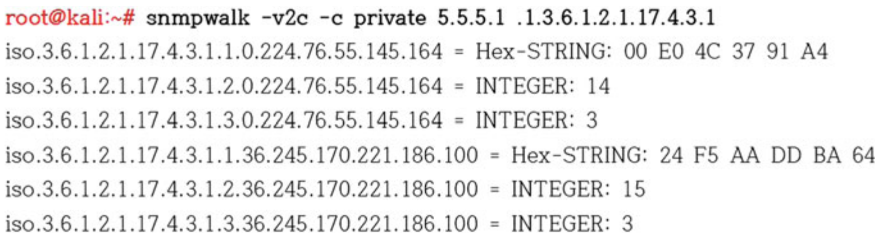


Fig. 3 Network testbed environment



```
root@kali:~# snmpwalk -v2c -c private 5.5.5.1 .1.3.6.1.4.1.9.9.46.1.6.1.1.14
iso.3.6.1.4.1.9.9.46.1.6.1.1.14.10114 = INTEGER: 1
iso.3.6.1.4.1.9.9.46.1.6.1.1.14.10115 = INTEGER: 1
```

Fig. 4 vlanTrunkPortDynam-icStatus from CISCO-VTP-MIB



```
root@kali:~# snmpwalk -v2c -c private 5.5.5.1 .1.3.6.1.2.1.17.4.3.1
iso.3.6.1.2.1.17.4.3.1.1.0.224.76.55.145.164 = Hex-STRING: 00 E0 4C 37 91 A4
iso.3.6.1.2.1.17.4.3.1.2.0.224.76.55.145.164 = INTEGER: 14
iso.3.6.1.2.1.17.4.3.1.3.0.224.76.55.145.164 = INTEGER: 3
iso.3.6.1.2.1.17.4.3.1.1.36.245.170.221.186.100 = Hex-STRING: 24 F5 AA DD BA 64
iso.3.6.1.2.1.17.4.3.1.2.36.245.170.221.186.100 = INTEGER: 15
iso.3.6.1.2.1.17.4.3.1.3.36.245.170.221.186.100 = INTEGER: 3
```

Fig. 5 *dot1dTpFdbTable* from BRIDGE-MIB

Second, we need to make sure that this interface works as a trunk port. Some Cisco switches manage CISCO-VTP-MIB, a proprietary MIB module from Cisco for the VLAN Trunk Protocol (VTP) and VLAN management [11]. In the MIB, *vlanTrunkPortDynamicStatus* (OID: 1.3.6.1.4.1.9.9.46.1.6.1.1.14) indicating whether the specified interface is either acting as a trunk or not. Its value can be one (Trunking) or two (notTrunking). In Fig. 4, we know that the two ports are trunk ports.

Third, bridging information of the switch is obtained from BRIDGE-MIB [12]. By querying all entries of *dot1dTpFdbTable* (OID: 1.3.6.1.2.1.17.4.3), the mapping information of MAC address and port can be obtained. The mapping information is called as ARP (Address Resolution Protocol) table, acting as a key role in layer-2 switching.

One could think that it can easily identify whether a PC is connected to the trunk port or not by using the trunk port information from the second step and the ARP table from the third step. Unfortunately, that is not simple. For example, although PC A is directly connected to the switch SA's trunk port in Fig. 3, there is no difference in ARP table information between PC A and B in Figs. 4 and 5.

Table 1 ARP population

Switch SA			Switch SB			Switch SC		
Port	Type	MAC	Port	Type	MAC	Port	Type	MAC
15	Trunk	24:F5:-	3	Trunk	24:F5:-	5	Trunk	24:F5:-
14	Trunk	00:EO:-	4	Trunk	00:EO:-	6	Access	00:EO:-

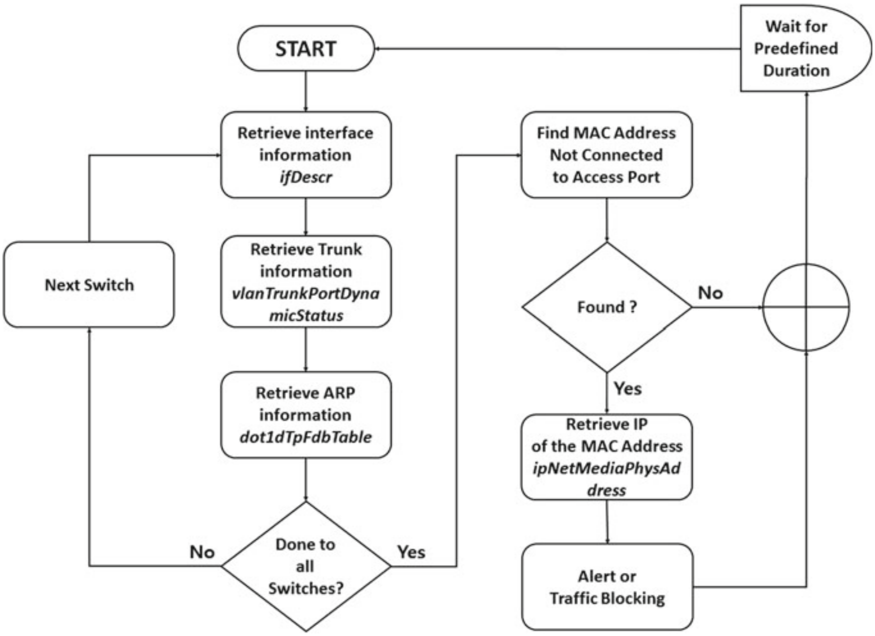

START
Retrieve interface information *ifDescr*
Retrieve Trunk information *vlanTrunkPortDynamicStatus*
Retrieve ARP information *dot1dTpFdbTable*
Done to all Switches?
No
Yes
Next Switch
Find MAC Address Not Connected to Access Port
Found ?
No
Yes
Retrieve IP of the MAC Address *ipNetMediaPhysAddress*
Alert or Traffic Blocking
Wait for Predefined Duration

Fig. 6 Searching algorithm for VLAN hopping attacking node

Our key idea to solve this problem is that we can utilize the following fact: one MAC address can be mapped to multiple trunk ports while it can be mapped to only one access port. For example, when PC A and C communicate, their MAC addresses appear in all the three switches as shown in Table 1. The MAC address of PC B is mapped to the access port of Switch SC, but that of PC A is never mapped to any access port. This means that PC A is directly connected to a trunk port of one of three switches.

Therefore, the fourth step is to iterate steps 1, 2, and 3 for all switches to retrieve the whole ARP mapping table and trunk port information to find out MAC addresses that are not mapped to access ports.

Finally, we query *ipNetMediaPhysAddress* (OID: 1.3.6.1.2.1.4.22.1.2) in order to find the IP address by the MAC address of the PC connected to the trunk port [13]. It is possible to prevent the threats early by notifying the administrator or automatically registering in the firewall to block the communication to the outside. Figure depicts shows the whole algorithm (Fig. 6).

5 Conclusion

In this paper, we proposed a method to effectively detect the occurrence of VLAN Hopping Attack in logical network separation environment using VLAN. Especially, since it uses SNMP which is applicable to many network components, this method will be effective because it can be deployed rather easily.

References

1. CISCO. Inter-Switch Link and IEEE 802.1Q Frame Format. https://www.cisco.com/c/en/us/support/docs/lan-switching/8021q/17056-741-4.html. Accessed 25 Aug 2006
2. IEEE Computer Society (2006) IEEE standard for local and metropolitan area networks—virtual bridged local area networks
3. SANS Institute (2016) Virtual LAN security weaknesses and countermeasures. https://www.sans.org/reading-room/whitepapers/networkdevs/virtual-lan-security-weaknesses-countermeasures-1090
4. David Hucaby. VLANs and Trunking. http://www.ciscopress.com/articles/article.asp?p=29803&seqNum=3. Accessed 25 Oct 2002
5. Cisco Networking Academy. Dynamic Trunking Protocol (3.2.3)>Cisco Networking Academy's Introduction to VLANs. http://www.ciscopress.com/articles/article.asp?p=2181837&seqNum=8. Accessed 7 Apr 2014
6. Convery S (2002) Hacking layer 2: fun with ethernet switches. https://www.blackhat.com/presentations/bh-usa-02/bh-us-02-convery-switches.pdf
7. Bhaij Y (2006) Layer 2 attacks & mitigation techniques. https://www.sanog.org/resources/sanog7/yusuf-L2-attack-mitigation.pdf
8. RFC 1157 (1990) A simple network management protocol (SNMP). https://www.ietf.org/rfc/rfc1157.txt
9. Rose M (1990) Management information base for network management of TCP/IP-based internets: MIB-II. https://tools.ietf.org/html/rfc1213
10. CNRS Grenoble, CRIC Homepage. http://cric.grenoble.cnrs.fr/Administrateurs/Outils/MIBS/?oid=1.3.6.1.2.1.2.2.1.2
11. CISCO-VTP-MIB. ftp://ftp.cisco.com/pub/mibs/v2/CISCO-VTP-MIB.my
12. RFC 4188. Definitions of managed objects for bridges. https://tools.ietf.org/html/rfc4188
13. CNRS Grenoble, CRIC Homepage. http://cric.grenoble.cnrs.fr/Administrateurs/Outils/MIBS/?oid=1.3.6.1.2.1.4.22.1.2

Improving Hidden Message Extraction Using LSB Steganalysis Techniques

Nikhil Mewalal and Wai Sze Leung

Abstract Increased awareness of the role of digital forensics in investigations has led to greater efforts being employed by users to conceal their data, possibly even using algorithms purposely designed to evade detection during steganalysis. A digital investigator seeking to ascertain whether some medium is indeed making use of steganography to hide pertinent evidence must therefore consider including other steganalysis techniques in their analysis in order to overcome the different steganographic strategies that may be used to evade detection. This paper investigates the design of a more comprehensive steganalysis tool that makes use of a series of statistical methods in conjunction with visual and forensic methods to detect messages hidden in images, specifically those hidden in PNG files using Least Significant Bit steganography. The study devises an appropriate combination of the techniques to generate a more effective and comprehensive steganalysis strategy for digital investigators attempting to detect hidden data.

Keywords Steganalysis · Steganography · Least significant bit

1 Introduction

A more comprehensive approach is required to conduct digital forensic investigations adequately when it comes to examining content for possible concealment through steganography.

Steganography is used as a means for communicating covertly in plain sight without arousing suspicion. This technique entails concealing a message inside an inconspicuous object in such a way that a casual observer is not able to differentiate between the original object (referred to as a cover object), and the object with the

N. Mewalal · W. S. Leung (✉)
University of Johannesburg, Johannesburg, South Africa
e-mail: wsleung@uj.ac.za

N. Mewalal
e-mail: 216086143@student.uj.ac.za

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_29

hidden message (also known as the stego image or steganogram). In the digital world, this cover object could be in the form of media, such as an image file, a video file, or even a music file [1].

Conversely, steganalysis refers to the complementary operation of detecting messages that have been hidden using steganography. During this process, the steganalysis algorithm is executed with the aim of interrogating the properties of an object which may be serving as the cover for hidden content with the aim of searching for any anomalies that could be deemed suspicious.

For digital forensics investigators, steganalysis is not without challenges. The existence of numerous steganography techniques (which can be dependent on the cover object file type), and the inability to extract the hidden message embedded in the first place can end up successfully frustrating steganalysis efforts.

In this paper, we propose the implementation of a more encompassing steganalysis tool that employs a greater variety of techniques to establish with greater certainty, whether a suspected file is a steganogram or simply a clean object. For the study, we will focus specifically on messages embedded into Portable Network Graphic (PNG) files. Our prototype will thus implement a series of Least Significant Bit (LSB) steganalysis techniques from varying domains to counter anti-detection attempts.

The rest of the paper is organized as follows: Sect. 2 briefly reviews some of the LSB steganalysis techniques to establish the different domains of LSB steganalysis techniques that exist. Section 3 then outlines a model of our more encompassing steganalysis solution. Section 4 presents the implementation details, along with the results of our prototype in Sect. 5, concluding the paper in Sect. 6.

2 Literature Review

2.1 Pairs of Values (Chi-Squared Attack)

This technique detects LSB embedding by using the histogram representation of the image. In a cover image, the gradient of the bars of the histogram can be noticeably smoother with bar heights varying throughout. In contrast, the bars on the histogram of a steganogram have neighbors that are roughly of equal height, producing a more rigid pattern [2]. Such an observation can be explained by how changes are applied within value pairs, resulting in an even distribution of 0 s and 1 s for each [2].

The success of histogram attacks on LSB steganography however relies on a fully-embedded image (where every single LSB is utilized to store a message).

2.2 Sample Pairs Analysis

When the image is only partially embedded along a pseudo-random path, examiners are no longer able to rely solely on a histogram of pixels while ignoring the dependency among neighboring pixels in natural images. For this reason, a more accurate and reliable method of detection that considers the spatial correlation within the image is used.

2.3 RS Analysis

RS Steganalysis relies on how LSB embedding operations work to determine whether an image contains a hidden message, as well as what the length of that image may be. This is achieved by identifying the presence of an imbalance in the cover file to establish the existence of hidden data [3, 4].

3 Model

Although the previous section has described several statistical techniques that could be employed for uncovering the presence of steganography, the use of just one domain area, such as statistical steganalysis, may not yield sufficiently thorough results. To address the shortcomings of a single steganalysis approach, we consider the application of several analysis models, which when used in conjunction with each other, provide a complementary set of techniques to test for most characteristics that reveal the presence of hidden messages.

3.1 Visual Models

LSB Amplification. This model aims to enhance the luminosity of the image in a manner that will end up removing all the parts of the image that is blocking the message. The human eye will then be able to distinguish whether there is a hidden message present in this message [5].

Difference and Neighborhood Histogram. Whilst this model may technically fall under a statistical method, it has been included as a visual one as it involves the production of a visual representation of the histogram which the user should ideally interpret. The neighboring bars in the histogram of a steganogram, for example, will exhibit similar heights [6].

3.2 Statistical Models

Chi-Square Attack. The approach to this form of steganalysis is to compare the expected frequency distribution in suspected stego images with a sample distribution observed in the possibly changed carrier image. Such an approach will however require an expected frequency distribution of the original cover image. Realistically, it would not be feasible to have a database of every possible image out there.

This shortcoming can however be addressed by estimating the theoretical expected frequency distribution [5].

Sample Pairs Analysis. This approach focuses on the analysis of transitions (such as slight color changes) in adjacent vertical or horizontal pixel pairs that are often imperceptible to the human eye [7]. An example is changing a pixel which is white (represented by FF FF FF) to FF FF FE.

3.3 Forensic Models

The model behind this attack is a custom one based on how LSB steganography works, reverse-engineering the process to extract the embedded message. The model aims to iterate through pixels in the domain search space and extract the least significant bit for that image. These values are then collected and concatenated in the end to produce a byte stream. This byte stream will be cast into a string value and then the result will be presented to the user.

4 Experimentation and Results

4.1 Preparation of Image Files

Images from the VOC 2005 Database: Dataset 1, as provided by the University of Oxford, were used for assessing our prototype implementation. The set contains 1578 images of categories motorbikes, bicycles, people, and cars in arbitrary poses [8]. From these, 26 images were selected and each processed to produce three different versions of the file. Details of the three versions are presented in Table 1.

Table 1 Preparation for each of the 26 images used for the experiment

Version	Embedded message (using LSB steganography)	Size (kb)
Clean image	None	N/A
Medium-length steganogram	US Constitution (pure text form)	26
Long-length steganogram	Lewis Carroll’s Alice in Wonderland (pure text form)	110

4.2 Steganographic Process

To embed the images with the medium- and long-length text messages described in Table 1, a LSB steganographic program developed for a separate, previous project was used. This program was tested extensively and verified as functioning correctly.

For each of the 26 images processed by the LSB steganographic program, two additional files with embedded messages were created.

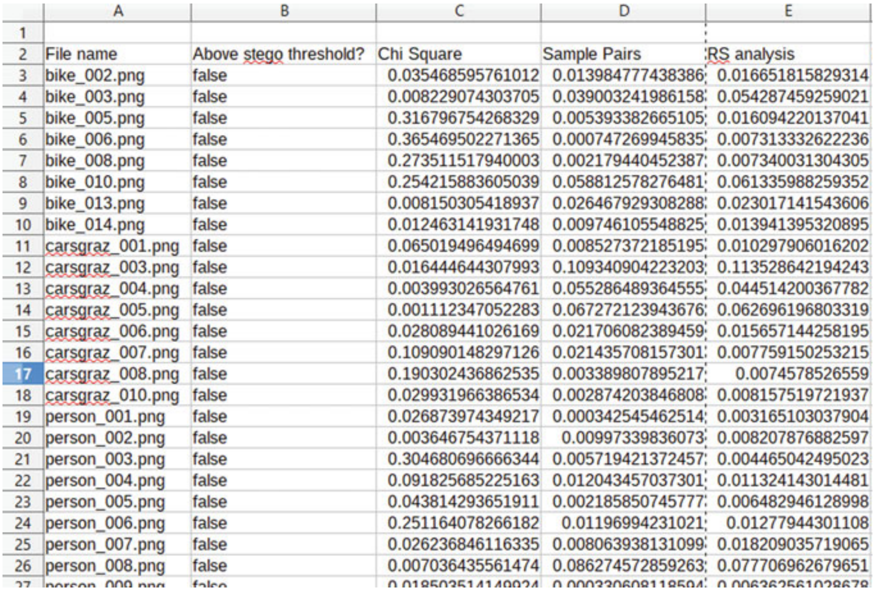
4.3 Testing of Statistics Methods

The implementation of the statistical algorithms incorporated in the model (Chi-Squared Attack, Sample Pairs Analysis, and RS Analysis) was assessed first. All 78 images prepared were run through the three statistical steganalysis techniques to obtain output in the form of percentages that suggested the possibility of steganography being embedded in the image. As seen in the results displayed in Figs. 1, 2 and 3, an additional threshold field was added to aggregate the results from the three statistical algorithms and reach a consensus-based outcome.

While the statistical algorithms were mostly accurate, it struggled to produce the correct result for several instances where the hidden message embedded was of medium-length.

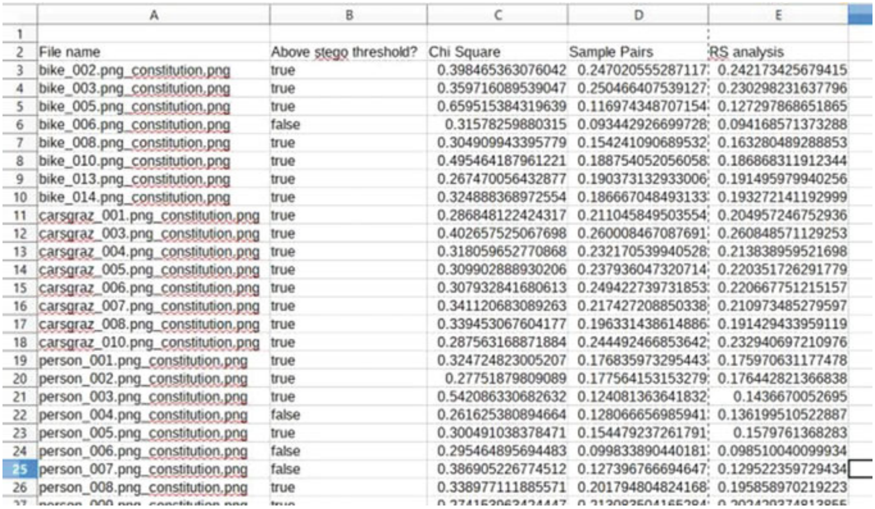
4.4 Testing of Visual Methods

Since visual methods require the input of a visual confirmation from a human user, the setup of our tests centered on LSB Amplification, Neighborhood Histogram, (and later the Forensic Steganalysis) are modified to focus on a single image set instead. Table 2 below shows the file sizes of each of the images.



	A	B	C	D	E
1					
2	File name	Above stego threshold?	Chi Square	Sample Pairs	RS analysis
3	bike_002.png	false	0.035468595761012	0.013984777438386	0.016651815829314
4	bike_003.png	false	0.008229074303705	0.039003241986158	0.054287459259021
5	bike_005.png	false	0.316796754268329	0.005393382665105	0.016094220137041
6	bike_006.png	false	0.365469502271365	0.000747269945835	0.007313332622236
7	bike_008.png	false	0.273511517940003	0.002179440452387	0.007340031304305
8	bike_010.png	false	0.254215883605039	0.058812578276481	0.061335988259352
9	bike_013.png	false	0.008150305418937	0.026467929308288	0.023017141543606
10	bike_014.png	false	0.012463141931748	0.009746105548825	0.013941395320895
11	carsgraz_001.png	false	0.065019496494699	0.008527372185195	0.010297906016202
12	carsgraz_003.png	false	0.016444644307993	0.109340904223203	0.113528642194243
13	carsgraz_004.png	false	0.003993026564761	0.055286489364555	0.044514200367782
14	carsgraz_005.png	false	0.001112347052283	0.067272123943676	0.062696196803319
15	carsgraz_006.png	false	0.028089441026169	0.021706082389459	0.015657144258195
16	carsgraz_007.png	false	0.109090148297126	0.021435708157301	0.007759150253215
17	carsgraz_008.png	false	0.190302436862535	0.003389807895217	0.0074578526559
18	carsgraz_010.png	false	0.029931966386534	0.002874203846808	0.008157519721937
19	person_001.png	false	0.026873974349217	0.000342545462514	0.003165103037904
20	person_002.png	false	0.003646754371118	0.00997339836073	0.008207876882597
21	person_003.png	false	0.304680696666344	0.005719421372457	0.004465042495023
22	person_004.png	false	0.091825685225163	0.012043457037301	0.011324143014481
23	person_005.png	false	0.043814293651911	0.002185850745777	0.006482946128998
24	person_006.png	false	0.251164078266182	0.01196994231021	0.01277944301108
25	person_007.png	false	0.026236846116335	0.008063938131099	0.018209035719065
26	person_008.png	false	0.007036435561474	0.086274572859263	0.077706962679651
27	person_009.png	false	0.018503514149924	0.000330608118594	0.006362561028678

Fig. 1 Results of statistical algorithms against cover images

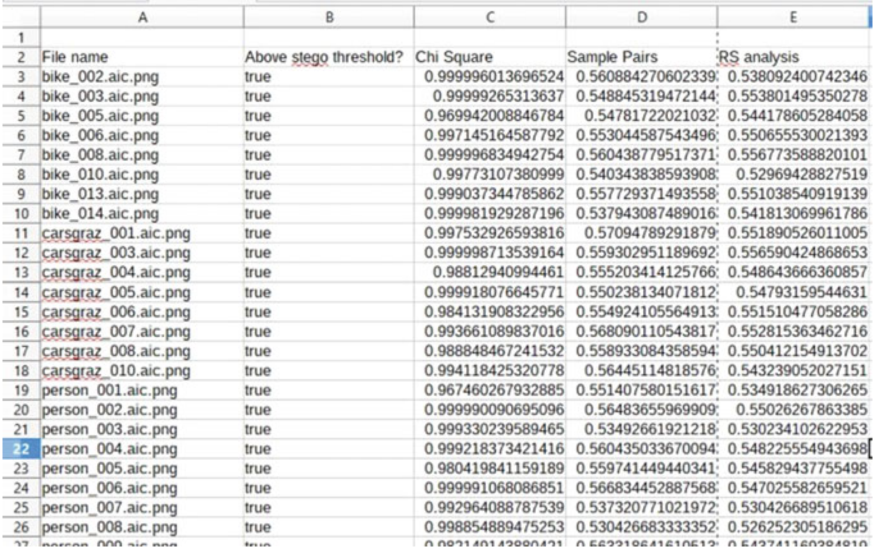


	A	B	C	D	E
1					
2	File name	Above stego threshold?	Chi Square	Sample Pairs	RS analysis
3	bike_002.png_constitution.png	true	0.398465363076042	0.247020555287117	0.242173425679415
4	bike_003.png_constitution.png	true	0.359716089539047	0.250466407539127	0.230298231637796
5	bike_005.png_constitution.png	true	0.659515384319639	0.116974348707154	0.127297868651865
6	bike_006.png_constitution.png	false	0.31578259880315	0.093442926699728	0.094168571373288
7	bike_008.png_constitution.png	true	0.304909943395779	0.154241090689532	0.163280489288853
8	bike_010.png_constitution.png	true	0.495464187961221	0.188754052056058	0.186868311912344
9	bike_013.png_constitution.png	true	0.267470056432877	0.190373132933006	0.191495979940256
10	bike_014.png_constitution.png	true	0.324888368972554	0.186667048493133	0.193272141192999
11	carsgraz_001.png_constitution.png	true	0.286848122424317	0.211045849503554	0.204957246752936
12	carsgraz_003.png_constitution.png	true	0.402657525067698	0.260008467087691	0.260848571129253
13	carsgraz_004.png_constitution.png	true	0.318059652770868	0.232170539940528	0.213838959521698
14	carsgraz_005.png_constitution.png	true	0.309902888930206	0.237936047320714	0.220351726291779
15	carsgraz_006.png_constitution.png	true	0.307932841680613	0.249422739731853	0.220667751215157
16	carsgraz_007.png_constitution.png	true	0.341120683089263	0.217427208850338	0.210973485279597
17	carsgraz_008.png_constitution.png	true	0.339453067604177	0.196331438614886	0.191429433959119
18	carsgraz_010.png_constitution.png	true	0.287563168871884	0.244492466853642	0.232940697210976
19	person_001.png_constitution.png	true	0.324724823005207	0.176835973295443	0.175970631177478
20	person_002.png_constitution.png	true	0.27751879809089	0.177564153153279	0.176442821366838
21	person_003.png_constitution.png	true	0.542086330682632	0.124081363641832	0.1436670052695
22	person_004.png_constitution.png	false	0.261625380894664	0.128066656985941	0.136199510522887
23	person_005.png_constitution.png	true	0.300491038378471	0.154479237261791	0.1579761368283
24	person_006.png_constitution.png	false	0.295464895694483	0.099833890440181	0.098510040099934
25	person_007.png_constitution.png	false	0.386905226774512	0.127396766694647	0.129522359729434
26	person_008.png_constitution.png	true	0.338977111885571	0.201794804824168	0.195858970219223
27	person_009.png_constitution.png	true	0.274153063424447	0.213083504165284	0.202420374813855

Fig. 2 Results of statistical algorithms against stego images (medium-length text)

4.5 Testing of Forensic Method

Because the clean images would have nothing to extract, this test focuses on recovering the hidden text from the stego images that were embedded with the medium-

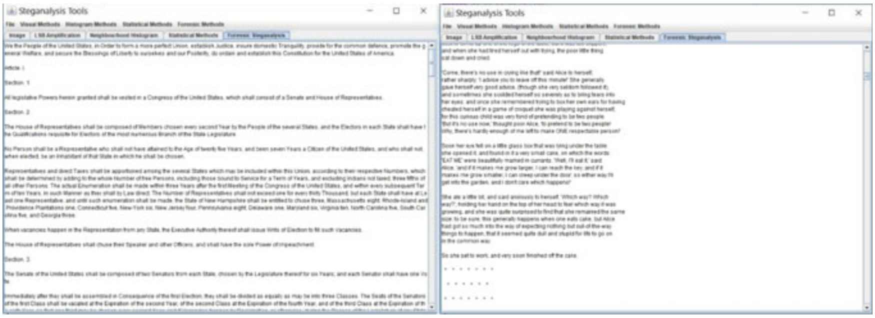


	A	B	C	D	E
1					
2	File name	Above stego threshold?	Chi Square	Sample Pairs	RS analysis
3	bike_002.aic.png	true	0.999996013696524	0.560884270602339	0.538092400742346
4	bike_003.aic.png	true	0.99999265313637	0.548845319472144	0.553801495350278
5	bike_005.aic.png	true	0.969942008846784	0.54781722021032	0.544178605284058
6	bike_006.aic.png	true	0.997145164587792	0.553044587543496	0.550655530021393
7	bike_008.aic.png	true	0.999996834942754	0.560438779517371	0.556773588820101
8	bike_010.aic.png	true	0.99773107380999	0.540343838593908	0.52969428827519
9	bike_013.aic.png	true	0.999037344785862	0.557729371493558	0.551038540919139
10	bike_014.aic.png	true	0.999981929287196	0.537943087489016	0.541813069961786
11	carsgraz_001.aic.png	true	0.997532926593816	0.57094789291879	0.551890526011005
12	carsgraz_003.aic.png	true	0.999998713539164	0.559302951189692	0.556590424868653
13	carsgraz_004.aic.png	true	0.98812940994461	0.555203414125766	0.548643666360857
14	carsgraz_005.aic.png	true	0.999918076645771	0.550238134071812	0.54793159544631
15	carsgraz_006.aic.png	true	0.984131908322956	0.554924105564913	0.551510477058286
16	carsgraz_007.aic.png	true	0.993661089837016	0.568090110543817	0.552815363462716
17	carsgraz_008.aic.png	true	0.988848467241532	0.558933084358594	0.550412154913702
18	carsgraz_010.aic.png	true	0.994118425320778	0.56445114818576	0.543239052027151
19	person_001.aic.png	true	0.967460267932885	0.551407580151617	0.534918627306265
20	person_002.aic.png	true	0.999990090695096	0.56483655969909	0.55026267863385
21	person_003.aic.png	true	0.999330239589465	0.53492661921218	0.530234102622953
22	person_004.aic.png	true	0.999218373421416	0.560435033670094	0.548225554943698
23	person_005.aic.png	true	0.980419841159189	0.559741449440341	0.545829437755498
24	person_006.aic.png	true	0.999991068086851	0.566834452887568	0.547025582659521
25	person_007.aic.png	true	0.992964088787539	0.537320771021972	0.530426689510618
26	person_008.aic.png	true	0.998854889475253	0.530426683333352	0.526252305186295

Fig. 3 Results of statistical algorithms against stego images (long-length text)

Table 2 File sizes of select image set used for experiment

Version	Size (kb)
Clean image	397
Stego image, embedded with the US Constitution (medium-length)	523
Stego image, embedded with Alice in Wonderland (long-length)	609


Steganalysis Tools
File Visual Methods Histogram Methods Statistical Methods Forensic Methods
Image LSB Amplification Neighborhood Histogram Statistical Methods Forensic Steganalysis
Steganalysis Tools
File Visual Methods Histogram Methods Statistical Methods Forensic Methods
Image LSB Amplification Neighborhood Histogram Statistical Methods Forensic Steganalysis

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.

Article. I.

Section. 1.

All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.

Section. 2.

The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

No Person shall be a Representative who shall not have attained to the Age of twenty five Years, and been seven Years a Citizen of the United States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence Plantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three.

When vacancies happen in the Representation from any State, the Executive Authority thereof shall issue Writs of Election to fill such Vacancies.

The House of Representatives shall chuse their Speaker and other Officers; and shall have the sole Power of Impeachment.

Section. 3.

The Senate of the United States shall be composed of two Senators from each State, chosen by the Legislature thereof for six Years; and each Senator shall have one Vote.

Immediately after they shall be assembled in Consequence of the first Election, they shall be divided as equally as may be into three Classes. The Seats of the Senators of the first Class shall be vacated at the Expiration of the second Year, of the second Class at the Expiration of the fourth Year, and of the third Class at the Expiration of th

and when she had tired herself out with trying, the poor little thing
sat down and cried.

'Come, there's no use in crying like that!' said Alice to herself,
rather sharply; 'I advise you to leave off this minute!' She generally
gave herself very good advice, (though she very seldom followed it),
and sometimes she scolded herself so severely as to bring tears into
her eyes; and once she remembered trying to box her own ears for having
cheated herself in a game of croquet she was playing against herself,
for this curious child was very fond of pretending to be two people.
'But it's no use now,' thought poor Alice, 'to pretend to be two people!
Why, there's hardly enough of me left to make ONE respectable person!'

Soon her eye fell on a little glass box that was lying under the table:
she opened it, and found in it a very small cake, on which the words
'EAT ME' were beautifully marked in currants. 'Well, I'll eat it,' said
Alice, 'and if it makes me grow larger, I can reach the key; and if it
makes me grow smaller, I can creep under the door; so either way I'll
get into the garden, and I don't care which happens!'

She ate a little bit, and said anxiously to herself, 'Which way? Which
way?', holding her hand on the top of her head to feel which way it was
growing, and she was quite surprised to find that she remained the same
size: to be sure, this generally happens when one eats cake, but Alice
had got so much into the way of expecting nothing but out-of-the-way
things to happen, that it seemed quite dull and stupid for life to go on
in the common way.

So she set to work, and very soon finished off the cake.

* * * * * * *

* * * * * *

* * * * * * *

Fig. 4 Results of text recovered from the two stego images using forensic steganalysis

and long-length messages. By applying the algorithm described in Sect. 3.3, we were able to recover the original text, as seen in Fig. 4.

5 Discussion of Results

5.1 *Statistical Techniques*

Medium-length messages hidden in stego images could at times result in the incorrect consensus that it was clean. A threshold was chosen such that any steganalyses producing probabilities higher than 20% would lead to the assumption that the image quite likely hides a message.

Images Embedded with Long-Length Messages. When it comes to images embedded with long-length messages, Chi-Squared tests perform incredibly well with the average certainty probability of the image containing steganography sitting at 99%. In contrast, both Sample Pairs and RS Analysis averaged at a probability of 55% and 54% respectively. Since the threshold is set at 20%, both algorithms can be seen to still yield positive, accurate results.

Images Embedded with Medium-Length Messages. As in the case with long-length embedded messages, Sample Pairs and RS Analysis performed similarly, detecting at an average of 19% and 18.6% respectively.

Seeing as the threshold value was 20%, these algorithms would have incorrectly classified many of the images as being clean, suggesting that such techniques do not perform well. In comparison, Chi-Squared tests yielded an average of 34.6%. Although this probability is also relatively low, the Chi-Squared test would still correctly classify the images due to the 20% threshold.

Such results serve to validate the appropriateness of the threshold level.

Clean Images. Once again, Sample Pairs and RS Analysis perform similarly. The average for steganography being present in clean images was around 2% for both. This average value is even lower at 1% using the Chi-Squared test. Overall, all 3 algorithms performed well as they did not incorrectly classify an image as being a stego object.

5.2 *LSB Amplification*

Clean Image. When using LSB Amplification on a clean image, the result is an image that resembles an old television set which is not tuned to a channel due to the completely random distribution of information. If this was the only image to sample as an observer, there would be little reason to doubt that there is anything suspicious about this image.

Image Embedded with Medium-Length Message. Because the image has been manipulated to embed a message, a definite pattern can be seen starting from the top left of the image, noticeable to the human eye. In the case of our selected image and

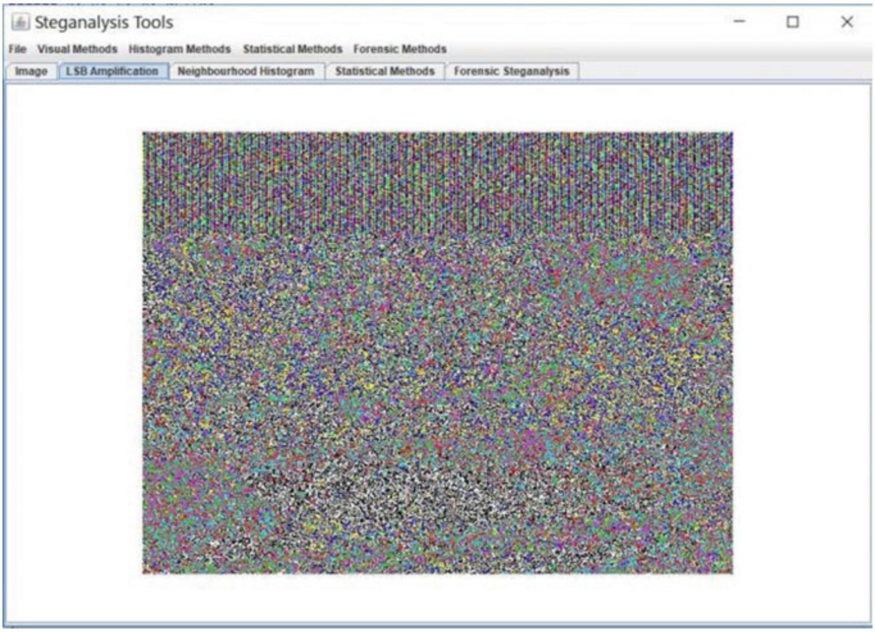

Steganalysis Tools
File Visual Methods Histogram Methods Statistical Methods Forensic Methods
Image LSB Amplification Neighbourhood Histogram Statistical Methods Forensic Steganalysis

Fig. 5 LSB Amplification with medium-length message

embedded image, we noticed that this pattern persisted for roughly the first 25% of the image's LSB enhancement (as seen in Fig. 5).

This observation can be attributed to the fact that given 397/523 * 100 = 75.9% (where the original file size is 397 kb while the medium length stego image is 523 kb), we have established how approximately 25% of the stego file contains hidden data.

Image Embedded with Long-Length Message. As noticed in the image embedded with a medium-length message, the extent at which noise is present in LSB Amplification in an image embedded with a message is dependent on the size of the message being hidden.

The noise level of the same image embedded with the longer Alice in Wonderland text is therefore much more pronounced. In such a case, a person asked to assess whether the image is embedded with a hidden message will have little reason to doubt that the image is indeed suspicious.

5.3 Neighborhood Histogram

Clean Image. As indicated by Westfeld, neighborhood histograms of images devoid of steganography generally have between 8 and 10 neighborhood colors [4]. In our test image, the neighborhood histogram produced 10 neighborhood colors, with all the bars of varying length.

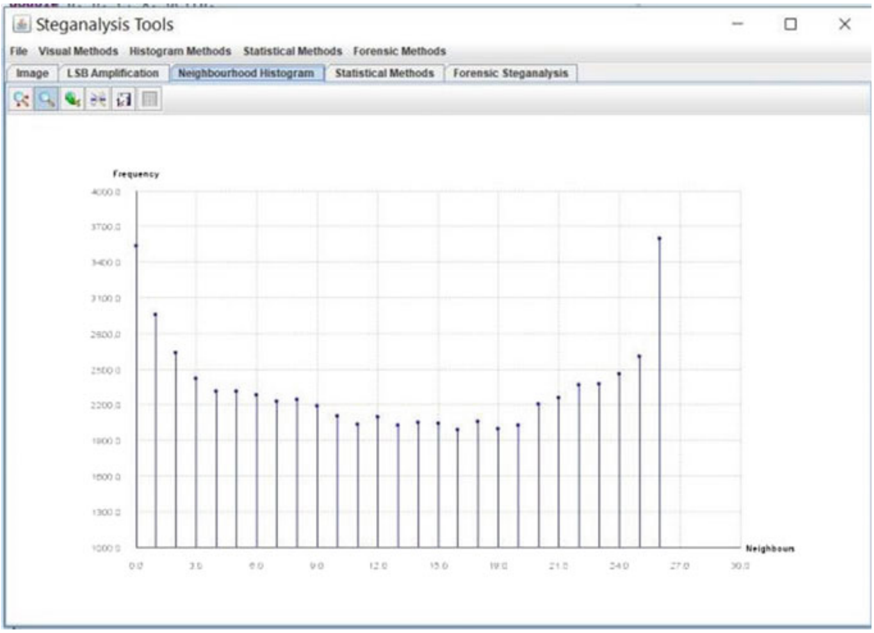


Fig. 6 Neighborhood histogram with long-length message

Image Embedded with Medium-Length Message. As established earlier in the LSB Amplification discussion, the embedded medium-length message takes up around 25% of the total file space. This observation accounts for the frequency of the neighborhood colors starting off very high and gradually lowering over time. The lower end of the histogram, which reflects typical behavior present in clean images, is the normal part of the image that does not contain any part of the concealed medium-length message.

Image Embedded with Long-Length Message. Due to the size of the message taking up almost the entire cover file space, the intensity of the frequency of neighborhood colors is consistently strong throughout the histogram. In this example (as seen in Fig. 6), most of the neighboring bars (as opposed to only some in the Medium-Length Message) are also of a similar height, making it easy for a human to visually discern that there is steganography embedded in the image.

6 Conclusion

Based on the experiments conducted, the following conclusions can be drawn:

- No one single statistical method can provide a solid and 100% detection rate—sometimes, the Chi-Square test will yield a better result than RS Analysis and sometimes, the opposite will be true.
- Combining results of the statistical methods leads to conclusions that are almost always accurate—this is especially true when a series of statistical techniques are combined, and a weighted average is used to draw the conclusion regarding the presence of steganography.
- Assessing each image for steganography using a combination of statistical methods is computationally taxing. On their own, each statistical test is already quite computationally expensive. This can be overcome by finding a mean of all the results to improve scanning time significantly. Specifically, the tool could start off with the tests known to be more accurate such as Chi-square and RS analysis. If both return good results, the steganalysis tool should not waste further resources by pursuing Primary Set and Sample Pairs Analysis.
- Employing visual techniques is a very strong way to identify suspicious images. However, this will need to be done on an image by image basis, which would ultimately prove non-feasible for investigators requiring results on a bulk set of images. For such purposes, investigators should revert to statistical steganalysis.
- The neighborhood histogram visual technique provided another very accurate approach to detecting steganography. The main concern here is that investigators employing this technique would need to have the knowledge to correctly interpret the result. It is however possible to programmatically develop the rules so that the interpretation of the histogram can be carried out by the computer instead.
- The initial results of the message extraction technique employed in the prototype proved to be positive, demonstrating how messages can be recovered from an image for forensic steganalysis purposes.

For further work, we anticipate testing our message recovery technique against other embedding techniques (other than LSB) to assess its validity.

Additional areas that will be focused on include the implementation of the programmatic interpretation of histograms and automating the decision-making concerning the optimal deployment of appropriate statistical methods for testing images for the presence of steganography.

References

1. Kessler GC, Hosmer C (2011) An overview of steganography. Adv Comput 83:51–107
2. Schaathun HG (2012) Machine learning in image steganalysis. Wiley, Norway
3. Bohme R (2010) Advanced statistical steganalysis. Springer

4. Manoharan S (2008) An empirical analysis of RS steganalysis. In: The third international conference on internet monitoring and protection internet monitoring and protection, pp 172–177
5. Westfeld A (2002) Detecting low embedding rates, Berlin
6. Zhang T, Ping X (2003) Reliable detection of LSB steganography based on the difference image histogram, Hong Kong
7. Shreelekshmi R, Wilsey M, Veni Madhavan C (2011) Improved LSB steganalysis based on analysis of adjacent pixel pairs. SIViP 7(5):811–816
8. Everingham M (2015) The PASCAL object recognition database collection. http://host.robots.ox.ac.uk/pascal/VOC/databases.html#VOC2005_1. Accessed 14 July 2017

Secrecy Analysis in the AF Mode Cooperative Communication System

Hsin-Ying Liang, Cheng-Ying Yang and Min-Shiang Hwang

Abstract Cooperative system is a tendency for the future communications because of its spatial diversity to improve the system performance. However, the security is a critical issue in the wireless application with a highly private request. Although the encryption schemes have been proposed to approach the secure purpose, those schemes need a lot of computing resource. It is not practical for the applications with a limited computing ability, such as IoT. According to Shannon theory of perfect secrecy, the security could be implemented on the physical layer. Based on the positive secrecy rate, the secure communication could be practical. This work concentrates on the theoretical solution to the secrecy rate in the AF mode cooperative communication system. Also, the numerical results are given. It shows the effects of eavesdropper could not affect the secure communication if the number of the eavesdropper is less than that of relays in the system. The appropriate relay assignment benefits the secure communication.

1 Introduction

Wireless communication networks play an important role in the smart city. It could provide a lot of advanced services. Although to access the wireless services is convenient, the degrading characteristics of the radio transmission are signal fading, mul-

H.-Y. Liang
Department of Info. & Comm. Engineering, Chaoyang University of Technology Taichung, Taichung, Taiwan

C.-Y. Yang
Department of Computer Science, University of Taipei, Taipei 100, Taiwan

M.-S. Hwang (✉)
Department of Computer Science & Information Engineering, Asia University, Taichung, Taiwan
e-mail: mshwang@asia.edu.tw

M.-S. Hwang
Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_30

tipath transmission, signal inferences, bandwidth limitation and so on [1, 2]. In order to combat the fading and increase the throughput, multiple-Input multiple-output (MIMO) that improves the system capacity, transmission speed and system performance have been realized as an effective scheme. However, with a high cost, MIMO could not be easily implemented because of physical size and power consumption [3]. Alternatively, the cooperative communication is an idea that the system makes the communication node helps each other to implement the communication process [4, 5]. The cooperative communication system provides a high throughput performance compared with multiple carrier modulation schemes and MIMO schemes. The destination user could transmit data with a spatial diversity by employing the relay stations.

By employing the relay station as the function of the antenna, the cooperative communication system increases the transmission data rate and provides a reliable channel capacity [6]. In the cooperative communication systems, the relay station functions with a character of spatial diversity. The relay station not only forwards the transmitted information but also process the received signal. It provides a high throughput performance. Hence, the cooperative communication is suitable to provide the multimedia services for the mobile devices. By employing the relay station as the virtual antenna, it increases the transmission data rate and provides a reliable channel capacity [7]. Besides, with a consideration of low cost, the cooperative communication system is a tendency in the future communications.

Three fundamental transmission modes are exited in the cooperative communications. One is Amplify-and-Forward (AF) mode. Another is Decode-and-Forward (DF) mode and the other is Compress and Forward (CF) mode [8]. With AF mode, the transmitted signal could be amplified and retransmit to the destination. It is easy with a low complexity. With a low complexity to implement in AF mode, it could be applied to those short distance transmissions [9]. However, security is an important factor in the wireless application with a highly private request. The critical issues of privacy and security have become increasingly important for the mobile users [10–14]. Especially, for banking and credit card transaction, security is an essential consideration for people to use the wireless application. Conventionally, the security depends on the cryptographic encryption at the application layer. RSA based asymmetric encryption and X.509 certifications were proposed [15]. Elliptic Curve Qu-Vanstone (ECQV) implicit certificate scheme and Elliptic Curve Diffie-Hellman (ECDH) key exchange scheme has been proposed [16–19]. Cryptographic encryption converts the meaningful information to be the apparent nonsense to avoid the eavesdroppers to release the transmitted information. However, the encryption algorithms are based on the assumption of limited computational capability at the eavesdroppers [20]. In addition, there exist the perfect encryption and the good protection schemes [21, 22]. For the real-time wireless application, e.g. the Internet of things (IoT) application, the ability of computing is limited, the security protection could be the limitation because of less computing resource [23–26]. Concerning of the secure communication, physical layer security could be practical [27–30].

Based on the maximum secrecy capacity, the analysis of AF mode cooperative communication system is proposed. It derives the relay assignment to maximize

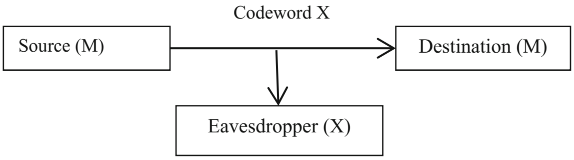


Fig. 1 A peer-to-peer communication with an eavesdropper

the overall secrecy rate. In the following section, the secure communication with Shannon theory is described. Also, the information in the cooperative system with AF mode is derived. Based on AF mode, the secure cooperative system model is proposed with the eavesdropper inside. In Sect. 4, the numerical results show the secrecy analysis for the cooperative system with an exhaustive search mechanism. The conclusion of this work is given in the final.

2 Secure Cooperative Communication System

Traditionally, a secure communication employs authorization and authentication schemes to control system access. Moreover, in a wireless system, an encryption is added to the existed protocol to protect the eavesdropper to catch the transmitted signal and decode to be the transmitted information. Critically, for the purpose of information security, it is reasonable to adopt the secure scheme at each layer. However, the security mechanism could be implemented with an efficient cost consideration. According to Shannon theory of perfect secrecy [31, 32], in Fig. 1, the security could be obtained if the entropy of the codeword is greater or equal to the transmitted information. It leads to realizing the uncertainty of the transmitted codeword must be at least as larger as the uncertainty of information [27], i.e. positive secrecy rate [32],

$$C_{s,d} = I_{s,d} - I_{s,e} \tag{1}$$

where $C_{s,d}$ is the secrecy rate (i.e. secrecy capacity) of transmission is defined as the mutual information difference between the mutual information from the source to the destination and that from the source to the eavesdropper. $I_{s,d}$ and $I_{s,e}$ denote the information between the source and the destination and the information between the source and the eavesdropper, respectively. Under AGWN, the information between the source to the destination1

$$I_{s,d} = \frac{1}{2}\log_2(1 + SNR_{s,d}) \tag{2}$$

where $SNR_{s,d}$ is defined as the signal power to noise ratio between the source and the destination.

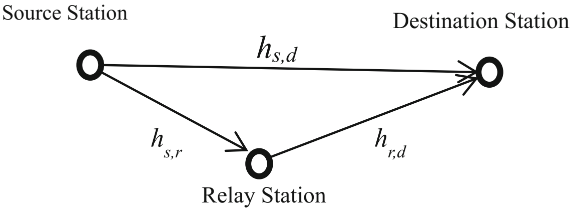


Fig. 2 A cooperative communication model

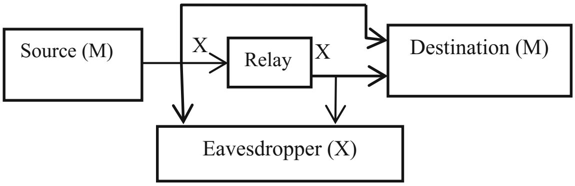


Fig. 3 Cooperative communication with an eavesdropper

Within Amplify-and-Forward (AF) mode, the transmitted signal could be amplified and retransmit to the destination. First, consider a single user in Fig. 2.
$h_{s,d}$ denotes the channel response between the source station to destination station. Similarly, $h_{s,r}$ and $h_{r,d}$ represent the channel response between the source station and the relay station and the channel response between the relay station and the destination station, respectively.

$$\begin{aligned} I_{AF} &= \frac{1}{2}\log_2(1 + SNR\mathrm{max}) \\ &= \frac{1}{2}\log_2\left(1 + \frac{P_s\left|h_{s,d}\right|^2}{N_0} + \frac{1}{N_0}\frac{P_sP_r\left|h_{s,r}\right|^2\left|h_{r,d}\right|^2}{P_s\left|h_{s,r}\right|^2 + P_r\left|h_{r,d}\right|^2 + N_0}\right) \end{aligned} \tag{3}$$

According to the above, the model of the secure communication system could be described in Fig. 3.

In Fig. 3, the eavesdropper locates at the end communication link. Let $h_{s,r}$ and $h_{r,d}$ denote as the channel response between the source station and the relay and the channel response between the relay and the destination, respectively. The mutual information between the source and the destination could be,

$$I_{s,d} = \frac{1}{2}\log_2\left(1 + \frac{P_s\left|h_{s,d}\right|^2}{N_0} + \frac{1}{N_0}\frac{P_sP_r\left|h_{s,r}\right|^2\left|h_{r,d}\right|^2}{P_s\left|h_{s,r}\right|^2 + P_r\left|h_{r,d}\right|^2 + N_0}\right) \tag{4}$$

where P_r is the signal power from the relay station and AWGN is with the variance N_0. Also, the mutual information between the source and the eavesdropper could be obtained [33].

$$I_{s,e} = \frac{1}{2}\log_2\left(1 + \frac{P_s\left|h_{s,e}\right|^2}{N_0} + \frac{1}{N_0}\frac{P_sP_r\left|h_{s,r}\right|^2\left|h_{r,e}\right|^2}{P_s\left|h_{s,r}\right|^2 + P_r\left|h_{r,e}\right|^2 + N_0}\right) \tag{5}$$

The eavesdropper could be the relay itself. Then, the mutual information between the source and the eavesdropper becomes

$$I_{s,e} = \frac{1}{2}\log_2\left(1 + \frac{P_s}{N_0}\left|h_{s,e}\right|^2\right) \tag{6}$$

The secrecy rate defined in Eq. (1). When the secrecy capacity is negative, the eavesdropper could intercept the transmitted information successfully. Hence, the condition for a secure communication, the secrecy rate, $C_{s,d}$, should be kept to be positive. The maximum of secrecy capacity $C_{s,d}$ could be reached by maximizing the mutual information between the source station and the destination station and minimizing the mutual information between the source station and the eavesdropper.

3 Secure Rate Analysis

To analyze the secrecy rate in the cooperative communications, initially, consider for the source station i transmits the information to the destination station $d(s_i)$ with the relay station r_i. Under AWGN channel, in AF mode, the secrecy capacity in the cooperative system becomes

$$C_{s_i,d(s_i)} = \max_{r=(r_1,r_2,\cdots r_k)\in R(s_1)\times R(s_2)\times\cdots\times R(s_k)}\left\{I_{s_i,d(s_i)} - I_{s_i,e}\right\} \tag{7}$$

The maximal mutual information achieved at the destination stations should consider the channel condition, under the multiple source station, multiple relay station and multiple destination station environments. Hence, the secure rate analysis for the secure cooperative communication could be developed based on the maximum mutual information between the information from the source station i to the corresponding destination station and the information from the source station i to the eavesdropper e. In the meantime, the secrecy capacity should be larger than 0 in Eq. (7) for the security purpose, i.e. the positive secrecy capacity $C_{si,d(si)}$. Hence, the limitation to this problem could become

$$C = \max\sum_{i=1}^{k}\sum_{j=1}^{m}\rho_{i,j}C_{s_i,d(s_i)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\rho_{i,j}\cdot\left\{\max\left(\left(I_{\mathrm{s,d(s)}}\right) - \left(I_{\mathrm{s,e}}\right)\right)\right\} \tag{8}$$

where $\rho_{i,j}$ is defined as the connection between the relay station i to the destination station j. In Eq. (8), for each destination, there is only one corresponding relay station connected to the destination station, $\rho_{i,j} = 1$ when there is a connection between relay station i to destination station j and $\rho_{i,j} = 0$ for other situations.

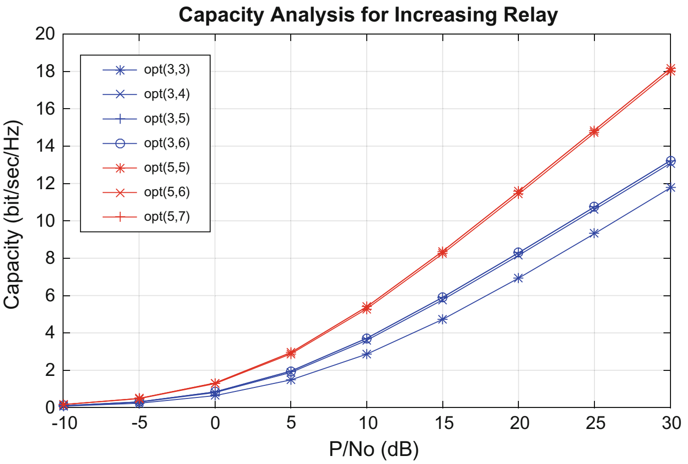


Fig. 4 System capacity of the cooperation system

4 Numerical Experimental Results

Based on the eavesdropper relay as an example, the numerical experiment calculates the combination for maximum value in Eq. (8) iteratively. The optimal mapping for the relay station and destination station could be found. It is with high complicated computing to implement. In Fig. 4, the symbol opt(m, n) denotes there are m resource stations and n relays, with an exhaustive algorithm to determine the optimal relay assignment, in the cooperation system. It shows the system capacity of the cooperation system. With the increasing the number of relays and the optimal relay assignment, the capacity will approach the maximum.

The secrecy capacity analysis of the systems with two eavesdroppers is given in Fig. 5. It shows if the system has extra relays to reduce the degrading effects of eavesdroppers. The secrecy rate is lightly affected if the number of the relay is larger than the number of eavesdroppers. However, the secrecy rate is seriously decreasing when the number relay is less. For the decreasing secret rate, the difference between the number of relay and number of the eavesdropper is the major degrading factor, but the ratio of relays and eavesdroppers is not. The relay mapping scheme could help to avoid the degrading effects of eavesdroppers and keep a positive secrecy capacity.

5 Conclusion

The cooperative system is a tendency for the future communications because of low cost and low complexity. In this work, the theoretical solution for a secret AF mode cooperative communication has been proposed. According to Shannon theory, with-

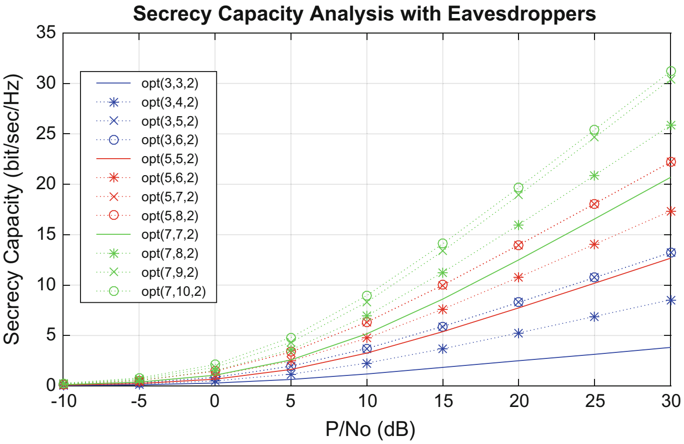


Fig. 5 Secrecy capacity of the system with two eavesdroppers

out the encryption, the physical layer could provide a secure transmission. Hence, the mutual information is a major concern in this work. Secrecy rate is the evaluation of the secure communication between the source to the destination. However, the security situation might be interrupted with the negative secrecy rate. In order to provide a secure communication, this work provides a relay mapping scheme to search the optimal assignment between the source and the relay. It begins to develop the theoretical limit for the secure communication. Then, based on the positive secrecy capacity, the maximum secrecy rate is approached with an appropriate relay assignment. In the numerical experiment, the results show if the number of eavesdroppers is less than that of the sources, the secure communication still could work with a suitable relay assignment.

Acknowledgements This research was partially supported by the Ministry Of Science and Technology, Taiwan (ROC), under contract no.: MOST 106-2221-E-845-001.

References

1. Singh R, Manu MS (2017) An energy efficient grid based static node deployment strategy for wireless sensor networks. Int J Electron Inf Eng 7(1):32–40
2. Rana A, Sharma D (2018) Mobile ad-hoc clustering using inclusive particle swarm optimization algorithm. Int J Electron Inf Eng 8(1):1–8
3. Chen X, Lei L, Zhang H, Yuen C (2015) Large-scale MIMO relaying techniques for physical layer security: AF or DF? IEEE Trans Wirel Commun 14(9):5135–5146
4. Nosratinia A, Hunter TE, Hedayat A (2004) Cooperative communication in wireless networks. IEEE Commun Mag 42(10):74–80

5. Agarwal S, Kansal T (2016) Congestion control schemes in ATM networks for ABR services: an overview. Int J Electron Inf Eng 5(2):57–67
6. Kramer G, Gastpar M, Gupta P (2005) Cooperative strategies and capacity theorems for relay networks. IEEE Trans Inf Theory 51(9):3037–3063
7. Wang Y, Noubir G (2013) Distributed cooperation and diversity for hybrid wireless networks. IEEE Trans Mobile Comput 12(3):596–608
8. Bletsas A, Shin H, Win MZ (2007) Outage analysis for cooperative communication with multiple amplify and forward relays. Electron Lett 43(6):51–52
9. Yang CY, Lin YS, Wen JH (2014) Greedy algorithm applied relay selection for cooperative communication systems in amplify-and-forward mode. J Electron Sci Technol 12(1):49–53
10. Choudhury H, Roychoudhury B, Saikia DKr (2016) Security extension for relaxed trust requirement in non3GPP access to the EPS hiten Choudhury. Int J Netw Secur 18(6):1041–1053
11. He L, Chen Y, Hu X, Qin Z (2017) An efficient and provably secure certificateless key insulated encryption with applications to mobile internet. Int J Netw Secur 19(6):940–949
12. Chiou SY, Ko WT, Lu EH (2018) A secure ECC-based mobile RFID mutual authentication protocol and its application. Int J Netw Secur 20(2):396–402
13. Arthi K, Reddy MC (2017) A secure and efficient privacy-preserving attribute matchmaking protocol for mobile social networks. Int J Netw Secur 19(3):421–429
14. Ngoc LT, Tu VT (2017) Whirlwind: a new method to attack routing protocol in mobile ad hoc network. Int J Netw Secur 19(5):832–838
15. Kothmayr T, Schmitt C, Hu W, Brünig M, Carle G (2013) DTLS based security and two-way authentication for the Internet of Things. Ad Hoc Netw 11(8):2710–2723
16. Hou G, Wang Z (2017) A robust and efficient remote authentication scheme from elliptic curve cryptosystem. Int J Netw Secur 19(6):904–911
17. Qian Q, Jia YL, Zhang R (2016) A lightweight RFID security protocol based on elliptic curve cryptography. Int J Netw Secur 18(2):354–361
18. Zhang X, Wang B, Wang W (2018) A new remote authentication scheme for anonymous users using elliptic curves cryptosystem. Int J Netw Secur 20(2):390–395
19. Han L, Xie Q, Liu W (2017) An improved biometric based authentication scheme with user anonymity using elliptic curve cryptosystem. Int J Netw Secur 19(3):469–478
20. Ng DWK, Lo ES, Schober R (2014) Robust beamforming for secure communication in systems with wireless information and power transfer. IEEE Trans Wirel Commun 13(8):4599–4615
21. Goswami S, Hoque N, Bhattacharyya DK, Kalita J (2017) An unsupervised method for detection of XSS attack. Int J Netw Secur 19(5):761–775
22. Srichavengsup W, San-Um W (2016) Data encryption scheme based on rules of cellular automata and chaotic map function for information security. Int J Netw Secur 18(6):1130–1142
23. Xu D, Wu Z, Wu Z, Zhang Q, Qin L, Zhou J (2015) Internet of Things: hotspot-based discovery service architecture with security mechanism. Int J Netw Secur 17(2):208–216
24. Porambage P, Schmitt C, Kumar P, Gurtov A, Ylianttila M (2014) Two-phase authentication protocol for wireless sensor networks in distributed IoT applications. In: Proceedings of the 2014 IEEE wireless communications and networking conference, Istanbul, Turkey, pp 2728–2733
25. Mayzaud A, Badonnel R, Chrisment I (2016) A taxonomy of attacks in RPL-based internet of things. Int J Netw Secur 18(3):459–473
26. Ma Y (2017) NFC communications-based mutual authentication scheme for the Internet of Things. Int J Netw Secur 19(4):631–638
27. Bloch M, Barros J (2011) Physical-layer security from information theory to security engineering. Cambridge
28. Tai WL, Chang YF (2017) Comments on a secure authentication scheme for IoT and cloud servers. Int J Netw Secur 19(4):648–651
29. Zou Y, Zhu J, Wang X, Leung V (2015) Improving physical-layer security in wireless communications using diversity techniques. IEEE Netw 29(1):42–48

30. Wang Q, Zhang H, Lyu Q, Wang X, Bao J (2018) A novel physical channel characteristics-based channel hopping scheme for jamming-resistant in wireless communication. Int J Netw Secur 20(3):439–446
31. Shannon CE (1949) Communication theory of secrecy systems. Bell Syst Tech J 29:656–715
32. Barros J, Rodrigues MR (2016) Secrecy capacity of wireless channels. In: Proceedings of 2006 IEEE international symposium on information theory, pp 356–360
33. Chen JS, Yang CY, Hwang MS (2017) The capacity analysis in the secure cooperative communication system. Int J Netw Secur 19(6):863–869

An Improved Hou-Wang's User Authentication Scheme

Min-Shiang Hwang, Hung-Wei Yang and Cheng-Ying Yang

Abstract It's easy to access Internet resources in the cloud environment. And it's important to protect the legal users' privacy and confidentiality. Recently, Hou and Wang proposed a robust and efficient user authentication scheme based on elliptic curve cryptosystem. Their scheme was practical and easy to implement. They claimed that their scheme could against off-line password guessing, DoS, server spoofing, replay, parallel session and impersonation attacks. In this article, we will show that Hou-Wang's scheme is vulnerable to the guessing attack with smart card. In this article, we also propose an improved Hou-Wang's user authentication scheme to withstand the vulnerability in their scheme.

Keywords Password · Smart card · User authentication

1 Introduction

It's easy to access Internet resources in the cloud environment. In order to protect the users could have the access right to obtain the resources provided by the remote server, the remote user authentication schemes were proposed [1–11]. Furthermore, it's also important to protect the legal users' privacy and confidentiality. To authenticate a user from Internet, many user authentication schemes had been proposed in past decades. Many schemes were applied a smart card to authenticate the legal users [12–21]. One of these schemes was developed for multi-servers [22–27]. One of these schemes

M.-S. Hwang · H.-W. Yang
Department of Computer Science & Information Engineering, Asia University, Taichung, Taiwan

M.-S. Hwang
Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

C.-Y. Yang (✉)
Department of Computer Science, University of Taipei, Taipei, Taiwan
e-mail: cyang@utaipei.edu.tw

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_31

was developed for biometrics [28–30]. One of these schemes was applied passwords for generating session key [31, 32].

In 2012, Li, Liu, and Wu proposed a secure remote user authentication to withstand the spoofing attack, forgery attack, and password guessing attack [33]. Unfortunately, Feng, Chao, and Hwang found the security of Li-Liu-Wu's scheme was vulnerable to password guessing attacks [34]. In 2012, Yoon et al. proposed an efficient remote user authentication scheme [35]. Unfortunately, Chen, Liang, and Hwang found their scheme is insecure to against the password guessing attack [36]. In 2014, Huang, Chang, Yu proposed a user authentication scheme which is based on timestamp [37]. Huabg et al. claimed their scheme could withstand the impersonated attack and more secure than other schemes. However, Feng, Liang, Hwang found that their scheme was vulnerable to the legal user's smart card and password guessing attack [38].

Recently, Hou and Wang proposed a robust and efficient user authentication scheme based on elliptic curve cryptosystem [39]. Hou-Wang's scheme is practical. They claimed that their scheme could against the off-line password guessing, DoS, spoofing, replay, parallel session, and impersonation attacks. In this article, we will show that Hou-Wang's scheme is vulnerable to the guessing attack with smart card. In this article, we also propose an improved Hou-Wang's user authentication scheme to withstand the vulnerability in their scheme.

2 Review of Hou-Wang Scheme

There are two main participants in Hou-Wang's scheme: a user U_i and server S [39]. We briefly describe Hou-Wang's scheme as follows.

The Registration Phase. In this registration phase, a new user (U_i) needs to apply to the server for as a legal user. After the phase, the server will make and issue a smart card for the new user (U_i). The smart card contains the following five parameters: $\{B_i$, H(), G, $E_k()$, and $D_k()\}$, here $B_i = E_{Ai}(H(x \| n_i) \| n_iG)$; $A_i = H(ID_i \| PW_i)$; where H() denotes a hash function; ID_i and PW_i denote an identity and password of the new user, respectively. x and n_i denote a server's master secret key and a random number for U_i, respectively. G denotes a public base point of elliptic curve; $E_k()$ and $D_k()$ denote an enciphering and deciphering algorithms with the secret key k, respectively. The server S maintains and keeps a registration table with two columns: $H(ID_i \oplus x)G$ and n_i.

The Login Phase. In this phase, when the user (U_i) wants to have the access right to obtain the resources provided by the remote server, U_i keys in his/her identity (ID_i) and password (PW_i) to the client devise with smart card. The smart card sends $\{C_i$, $D_i\}$ to the server S: $Ai = H(ID_i \| PW_i)$; $B_i = E_{Ai}(H(x \| n_i) \| n_iG)$; $H(x \| n_i) \| n_iG =$

$D_{Ai}(B_i)$; $C_i = t\,G$; $K_i = t\,Pub_s$; $D_i = E_{Ki}(ID_i \,||\, H(x \,||\, n_i))$, where t denotes a random nonce in Z_p*. Pubs is the server's public key, $Pub_s = x\,G$.

The Authentication and Session Key Exchange Phase. In this authentication and session key exchange phase, the server (S) verifies U_i as follows.

(1) After receiving $\{C_i, D_i\}$, the server calculates and obtains the deciphering key K_i, U_i, and $H(x \,||\, n_i)$ as follows: $K'_i = x\,C_i$; $ID'_i \,||\, H(x' \,||\, n'_i) = D_{K'i}(D_i)$. Next, S computes $H(ID'_i \oplus x)G$ and retrieves the random number ni of U_i from the registration table.

(2) S computes $H(x \,||\, n_i)$ and then verifies $H(x \,||\, n_i)$ is whether or not equal to $H(x' \,||\, n'_i)$. If it is not holds, S terminates this phase. Next, S sends $\{E_i, F_i\}$ to U_i, where $E_i = s\,G$; $F_i = s\,C_i + n_iG$, where s denotes a random nonce in Z_p*.

(3) The smart card checks E_i and F_i. The server also authenticates the legal user. Finally, the server and smart card share the session key $SK = stG$.

3 The Weakness and the Improved of Hou-Wang Scheme

In this section, we show the weakness of Hou-Wang's remote user authentication scheme [39]. The main weakness of Hou-Wang's scheme is that their scheme could not against the on-line password guessing attack with user's smart card (SC for short). A user U_i's smart card may be lost or stolen by an adversary. The adversary could try to guess the user's password.

(1) The adversary inserts the user U_i's smart card to his/her client device. Next, the adversary keys in the identity of the user U_i and guesses a password PW'_i.

(2) SC sends $\{C_i, D_i\}$ to the server S: $A'_i = H(ID_i \,||\, PW'_i)$; $B_i = E_{Ai}(H(x \,||\, n_i) \,||\, n_iG)$; $H'(x \,||\, n_i) \,||\, n'_iG = D_{A'i}(B_i)$; $C_i = t\,G$; $K_i = t\,Pub_s$; $D_i = E_{Ki}(ID_i \,||\, H'(x \,||\, n_i))$.

3) The server performs Steps (1) and (2) in the authentication and session key exchange phase to verify the user (adversary) legally. If the guessing password by the adversary is correct, the adversary will receive $\{E_i, F_i\}$ from the server. Otherwise, the adversary guesses the other password PW'_i and repeats Step (1).

In order to improve the weakness of Hou-Wang's remote user authentication scheme, we propose an improvement of Hou-Wang's scheme in this section. The password changing and the smart revocation phases are the same as that in Hou-Wang's scheme.

The Registration Phase. In this phase, a new user (U_i) needs to apply to the server for as a legal user. After the phase, the server will make and issue a smart card for U_i. The smart card contains $\{B_i$, H(), G, $E_k()$, and $D_k()\}$, where $B_i = E_{Ai}(H(x \,||\, n_i) \,||\, n_iG)$; $A_i = H(ID_i \,||\, PW_i)$. The server S maintains and keeps a registration table with three columns: $H(ID_i \oplus x)G$, ni, and counter (see Table 1). The counter is used to record the times of failing to login the server.

Table 1 The registration table

User's identity	Nonce	Counter
$H(ID_1 \oplus x)G$	n_1	0
$H(ID_2 \oplus x)G$	n_2	2
:	:	:
$H(ID_i \oplus x)G$	n_i	1
:	:	:
$H(ID_m \oplus x)G$	n_m	0

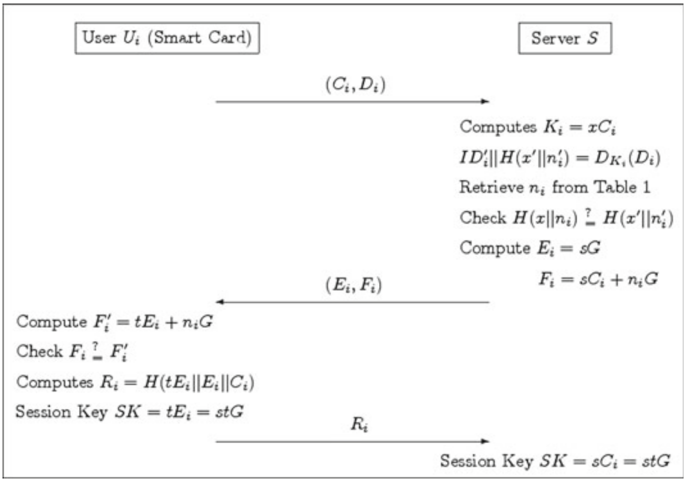

User U_i (Smart Card)
Server S
(C_i, D_i)
Computes $K_i = xC_i$
$ID'_i || H(x'||n'_i) = D_{K_i}(D_i)$
Retrieve n_i from Table 1
Check $H(x||n_i) \overset{?}{=} H(x'||n'_i)$
Compute $E_i = sG$
$F_i = sC_i + n_iG$
(E_i, F_i)
Compute $F'_i = tE_i + n_iG$
Check $F_i \overset{?}{=} F'_i$
Computes $R_i = H(tE_i||E_i||C_i)$
Session Key $SK = tE_i = stG$
R_i
Session Key $SK = sC_i = stG$

Fig. 1 The authentication and session key exchange phase of our scheme

The Login Phase. This phase is similar to that of Hou-Wang scheme. In this phase, when U_i wants to have the access right to obtain the resources provided by the remote server, U_i keys in his/her identity (ID_i) and inputs his/her password (PW_i) to the client devise with smart card. The smart card sends $\{C_i, D_i\}$ to the server S: $A_i = H(ID_i || PW_i)$; $H(x || n_i) || n_iG = D_{Ai}(B_i)$; $C_i = t\ G$; $K_i = t\ Pub_s$; $D_i = E_{Ki}(ID_i || H(x || n_i))$.

The Authentication and Session Key Exchange Phase. In this authentication and session key exchange phase, S verifies Ui as follows (see Fig. 1).

(1) After receiving $\{C_i, D_i\}$, the server calculates and obtains the deciphering key K_i, the U_i identity, and $H(x || n_i)$ as follows: $K'_i = x\ C_i$; $ID'_i || H(x' || n'_i) = DK'_i(D_i)$.
(2) S computes $H(ID'_i \oplus x)G$ and retrieves the random number n_i of U_i from Table 1.
(3) S computes $H(x || n_i)$ and then verifies $H(x || n_i)$ is whether or not equal to $H(x' || n'_i)$. If it is not holds, the server stops this procedure and adds 1 to the counter in Table 1. If the counter is greater than 3, the server removes the user's

information from registration table. The user needs to re-makes a registration for sharing the server's resource.

(4) The server S sends {E_i, F_i} to the user U_i, where E_i = s G; F_i = s C_i + n_i G, where s denotes a random nonce in Z_p*.

(5) The smart card computes F'_i = tE_i + n_i and then checks F'_i is whether or not equal to F_i. If it holds, computes and sends the verification message R_i to the server: $R_i = H(tE_i \parallel E_i \parallel C_i)$.

(6) The server computes $R'_i = H(sC_i \parallel E_i \parallel C_i)$ and checks R'_i whether equal to R_i. If it holds, S thus authenticates the legal user.

(7) The server and the smart card share the session key SK = stG.

Subsequent paragraphs, however, are indented.

4 Conclusions

In summary, we have shown that the weakness of Hou-Wang's remote user authentication scheme. Hou-Wang's scheme could not against the on-line password guessing attack with smart card. In this article, we also proposed an improvement of Hou-Wang's remote user authentication scheme to improve the weakness in Hou-Wang's scheme.

Acknowledgements This work was partially supported by the Ministry of Science and Technology, Taiwan, under grant MOST 106-2221-E-468-002.

References

1. Tsai CS, Lee CC, Hwang MS (2006) Password authentication schemes: current status and key issues. Int J Netw Secur 3:101–115
2. Yang CC, Chang TY, Hwang MS (2003) The security of the improvement on the methods for protecting password transmission. Informatica 14:551–558
3. Zhuang X, Chang CC, Wang ZH, Zhu Y (2014) A simple password authentication scheme based on geometric hashing function. Int J Netw Secur 16:271–277
4. Ling CH, Chao WY, Chen SM, Hwang MS (2015) Cryptanalysis of dynamic identity based on a remote user authentication scheme for a multi-server environment. In: Advances in engineering research, vol 15. Atlantis Press, pp 981–986
5. Liu Y, Chang CC, Chang SC (2017) An efficient and secure smart card based password authentication scheme. Int J Netw Secur 19(1):1–10
6. Liu CW, Tsai CY, Hwang MS (2017) Cryptanalysis of an efficient and secure smart card based password authentication scheme. In: Advances in intelligent systems and computing, recent developments in intelligent systems and interactive applications, vol 541. Springer, pp 188–193 (2017)
7. Wei J, Liu W, Hu X (2016) Secure and efficient smart card based remote user password authentication scheme. Int J Netw Secur 18(4):782–792

8. Tsai CY, Pan CS, Hwang MS (2017) An improved password authentication scheme for smart card. In: Advances in intelligent systems and computing, recent developments in intelligent systems and interactive applications, vol 541. Springer, pp 194–199
9. Thandra PK, Rajan J, Satya Murty SAV (2016) Cryptanalysis of an efficient password authentication scheme. Int J Netw Secur 18(2):362–368
10. Pan CS, Tsai CY, Tsaur SC, Hwang MS (2016) Cryptanalysis of an efficient password authentication scheme. In: The 3rd IEEE international conference on systems and informatics, Shaihai, pp 732–737
11. Pan HT, Pan, CS, Tsaur, SC, Hwang, MS (2017) Cryptanalysis of efficient dynamic ID based remote user authentication scheme in multi-server environment using smart card. In: 12th international conference on computational intelligence and security, Wuxi, China, pp 590–593
12. He D, Chen J, Hu J (2011) Weaknesses of a remote user password authentication scheme using smart card. Int J Netw Secur 13:58–60
13. Hwang MS, Chong SK, Chen TY (2000) Dos-resistant ID-based password authentication scheme using smart cards. J Syst Softw 83:163–172
14. Hwang MS, Li LH (2000) A new remote user authentication scheme using smart cards. IEEE Trans Consum Electron 46:28–30
15. Kumar M, Gupta MK, Kumari S (2011) An improved efficient remote password authentication scheme with smart card over insecure networks. Int J Netw Secur 13:167–177
16. Ramasamy R, Muniyandi AP (2012) An efficient password authentication scheme for smart card. Int J Netw Secur 14:180–186
17. Shen JJ, Lin CW, Hwang MS (2003) Security enhancement for the timestamp-based password authentication scheme using smart cards. Comput Secur 22:591–595
18. Shen JJ, Lin CW, Hwang MS (2003) A modified remote user authentication scheme using smart cards. IEEE Trans Consum Electron 49:414–416
19. Tang H, Liu X, Jiang L (2013) A robust and efficient timestamp-based remote user authentication scheme with smart card lost attack resistance. Int J Netw Secur 15:446–454
20. Yang L, Ma JF, Jiang Q (2012) Mutual authentication scheme with smart cards and password under trusted computing. Int J Netw Secur 14:156–163
21. Ghosh D, Li C, Yang C (2018) A lightweight authentication protocol in smart grid. Int J Netw Secur 20(3):414–422
22. Feng TH, Ling CH, Hwang MS (2014) Cryptanalysis of Tan's improvement on a password authentication scheme for multi-server environments. Int J Netw Secur 16:318–321
23. He D, Zhao W, Wu S (2013) Security analysis of a dynamic id-based authentication scheme for multi-server environment using smart cards. Int J Netw Secur 15:282–292
24. Li LH, Lin IC, Hwang MS (2001) A remote password authentication scheme for multi-server architecture using neural networks. IEEE Trans Neural Netw 12:1498–1504
25. Lin IC, Hwang MS, Li LH (2003) A new remote user authentication scheme for multi-server architecture. Futur Gener Comput Syst 19:13–22
26. Amin R (2016) Cryptanalysis and efficient dynamic ID based remote user authentication scheme in multi-server environment using smart card. Int J Netw Secur 18(1):172–181
27. Mohan NBM, Chakravarthy ASN, Ravindranath C (2018) Cryptanalysis of design and analysis of a provably secure multi-server authentication scheme. Int J Netw Secur 20(2):217–224
28. Li CT, Hwang MS (2010) An online biometrics-based secret sharing scheme for multiparty cryptosystem using smart cards. Int J Innov Comput Inf Control 6:2181–2188
29. Li CT, Hwang MS (2010) An efficient biometrics-based remote user authentication scheme using smart cards. J Netw Comput Appl 33:1–5
30. Prakash A (2014) A biometric approach for continuous user authentication by fusing hard and soft traits. Int J Netw Secur 16:65–70
31. Zhu H, Zhang Y (2017) An improved two-party password-authenticated key agreement protocol with privacy protection based on chaotic maps. Int J Netw Secur 19(4):487–497
32. Wu M, Chen J, Wang R (2017) An enhanced anonymous password-based authenticated key agreement scheme with formal proof. Int J Netw Secur 19(5):785–793

33. Li J, Liu S, Wu S (2012) Cryptanalysis and improvement of a YS-like user authentication scheme. Int J Digit Conten Technol Appl 7(1):828–836
34. Feng TH, Chao WY, Hwang MS (2014) Cryptanalysis and improvement of the Li-Liu-Wu user authentication scheme. In: International conference on future communication technology and engineering, Shenzhen, China, pp 103–106
35. Yoon EJ, Kim SH, Yoo KY (2012) A security enhanced remote user authentication scheme using smart cards. Int J Innov Comput, Inf Control 8(5):3661–3675
36. Chen TY, Ling CH, Hwang MS (2014) Weaknesses of the Yoon-Kim-Yoo remote user authentication scheme using smart cards. In: IEEE workshop on electronics, computer and applications, Ottawa, Canada, pp 771–774
37. Huang HF, Chang HW, Yu PK (2014) Enhancement of timestamp-based user authentication scheme with smart card. Int J Netw Secur 16:463–467
38. Feng TH, Ling CH, Hwang MS (2014) An improved timestamp-based user authentication scheme with smart card. In: The 2nd congress on computer science and application, Sanya, China, pp 111–117 (2014)
39. Hou G, Wang Z (2017) A robust and efficient remote authentication scheme from elliptic curve cryptosystem. Int J Netw Secur 19(6):904–911

Developing a Testing Framework for Intrusion Detection Algorithms Using Software Defined Networking

Anton Miguel Suba, Kurt Vincent Bautista, Julio Carlos Tomas Ledesma and William Emmanuel Yu

Abstract Software defined networking (SDN) is an emerging type of network technology that aims to make the network flexible and adaptable. This paper presents a study that explores the creation of a testing framework for intrusion detection systems (IDS) created using SDN. IDSes created using SDN have a distinct flexibility and configurability that current network security do not have. While there have been a number of network security software created using SDN, there is a lack of a way to easily test these software and show results. This study aimed to create a tool that would test these systems and allow for easy generation of network topologies, training of machine learning models, and swapping of test scripts. The methodology entails the creation of the testing framework to test IDSes in an intuitive and user-friendly way, then using a machine learning IDS created using SDN to test the effectiveness of the testing framework. The results of the experiment show that the framework was able to successfully test an IDS, and give accurate results.

Keywords Intrusion detection system · Machine learning · Mininet
Software defined networking · Testing framework

A. M. Suba (✉) · K. V. Bautista · J. C. T. Ledesma · W. E. Yu
Ateneo de Manila University, Katipunan Avenue, 1108 Quezon, Metro Manila, Philippines
e-mail: anton.suba@obf.ateneo.edu

K. V. Bautista
e-mail: kurt.bautista@obf.ateneo.edu

J. C. T. Ledesma
e-mail: julio.ledesma@obf.ateneo.edu

W. E. Yu
e-mail: wyu@ateneo.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_32

1 Introduction

Network security is of paramount importance in a data-centric world. Organizations and individuals need efficient and secure methods for delivery and storage of potentially sensitive data. Most of the networks that are used today are implemented with specific hardware in mind. Modification of said hardware to adopt new functions and protocols is often impossible, while intruders can keep finding and exploiting network vulnerabilities. With the ever-changing landscape of network security, traditional networks are left susceptible to new forms of attacks.

To combat these attacks in a cost-effective and adaptable manner, SDN has recently emerged as one of the leading solutions to address the limitations of currently existing networks. SDN can be defined as "the physical separation of the network control plane from the forwarding plane, and where a control plane controls several devices" [1]. This decoupling allows the network control to be directly programmable by the user and able to adapt to new kinds of attacks.

While SDN has been proven to be able to bolster network security, the field is still relatively new, and lacks a way to evaluate and test these new network technologies. In this study, a framework was created in order to test and evaluate IDSes created using SDN. Our objectives are as follows: (1) To create a testing framework that can intuitively, and accurately test IDSes. (2) To provide tools to the user that will assist in the creation of machine learning IDSes. Our research methodology focuses on looking at past implementations, and past tests of network security devices that were created using SDN in order to create a suitable testing framework to test machine learning IDSes.

2 Prior Research

There have been a number of previous studies that use SDN in order to create network security software, as well as multiple researches that aim to create a valid metric in order to evaluate intrusion detection systems. This study examined the methodologies and the research of these experiments in order to create the framework. We also heavily modified some parts of past researches.

2.1 Security Software Using SDN

A study by Guevara et al. [2] focused on creating an IDS using SDN. They specifically tried to detect and stop the Black Nurse Distributed Denial of Service (DDoS) attack. They enhanced the intrusion detection system by distributing the detection process to the various host-connected switches within the network. Their experiment was a success, having a much higher performance during an attack, while only having a slight loss in performance when no attacks are occurring. This study took the idea of using a distributed topology, and heavily modified it in order to fit our study. This study also adopted the switches they created.

Another paper written by Peña focused on creating a distributed firewall using SDN [3]. Peña aimed to create a firewall that would maximize the features of SDN. He took advantage of the centralized control, and the distributed properties of each network device. We used Peña's idea of a distributed topology as a guide for the network topology included in the framework.

A paper written by Mell [4], talks about the difficulty of testing IDSes, talking about how the difference in configurations can make it difficult to compare IDSes, as well as the difficulties in collecting test scripts. Finally, she outlines four approaches to testing IDSes with background data. These four approaches were taken into account during the creation of the IDS, as well as the validation tests.

2.2 Machine Learning Intrusion Detection Systems

A study by Dongre and Wankhade successfully creates a machine learning IDS using an ensemble approach, specifically a boosting tree [5]. Their research involved finding a solution to the common problem of databases. Whenever structural changes are made to the database, the database exposes itself to risks. IDSes are created to protect them, but traditional IDSes can still be improved as they are still quite inefficient, and they still have a need for human administrators for unfamiliar attacks. This research would use the Adaptive Boosting(Adaboost) algorithm for the machine learning IDS that would be used for testing.

2.3 Metrics for Evaluating Intrusion Detection Systems

A paper by Kumar titled Evaluation Metrics for Intrusion Detection Systems—A Study. Talks about the various metrics in evaluating intrusion detection systems [6]. He lists the three classes of metrics namely threshold, ranking, and probability metrics. Of these, the research chose to use ranking metrics such as False positive rate, Detection rate, and Precision. Kumar's paper also shows the usage of a confusion matrix to represent the results of a test. Our research took note of the various metrics in testing the IDSes and we decided to use it as a basis for the framework's metric logger module. The framework used the confusion matrix, as well as the precision and recall scores as the metrics by which the framework would evaluate IDses.

Another paper by Munaiah et al. titled "Are Intrusion Detection Studies Evaluated Consistently? A Systematic Literature Review" [7] aimed to find commonly used evaluation metrics for IDSes and their related literature to guide other researchers. It was shown that the Confusion Matrix as well as the Precision and Recall scores were almost perfect in terms of testing the effectiveness of an IDS. Because the testing

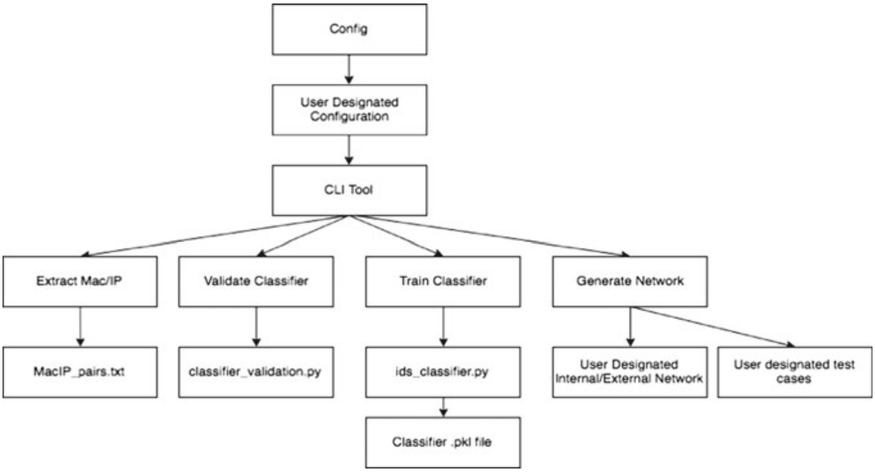


Fig. 1 Overview of the config and the CLI Tool

framework of this study only tests the effectiveness of the IDS, the metrics chosen were from the effectiveness category.

3 Methodology

The framework is a collection of user-friendly tools that allows users to easily switch between machine learning algorithms and models, test scripts, topologies, and the IDSes used. We first identified specific pain points involved in the process of building and testing IDSes that could be improved. Specifically: model training and validation, network generation and execution of test scenarios. Specific modules were then implemented to address such issues. In order to test the validity of the framework, a machine learning model was trained, cross-validated, and then used as the core classifier of an IDS. Custom test scenarios were then scripted using the framework testing module and executed against the IDS.

3.1 System Overview

Mininet is used for rapid prototyping and network simulation. POX controller would allow us to write controllers using the Python programming language. POX is a controller for the switches made with Mininet. Figure 1 above shows how Mininet, and POX are connected, as well as the other components required (Fig. 2).

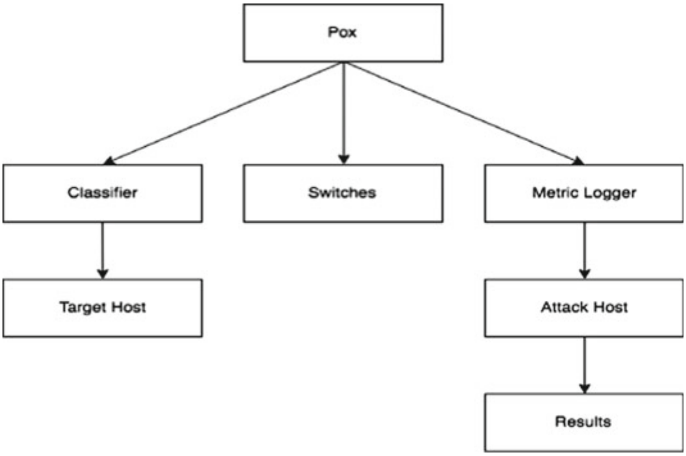


Fig. 2 Overview of the pox controller. Showing the requirements and the outputs

As was previously said, Mininet is responsible for simulating the network. Using MAC address and IP address pairs obtained from the PCAP files, the network topology is generated through a Python script. This component contains the topology for the network, the configuration, as well as the test cases. Test cases are Python modules that are executed by the framework in sequence.

3.2 Framework Architecture

The testing framework was designed to be robust and configurable. The framework has four primary functions: model training, model validation, network generation and network-specific variable extraction, contained in their own Python modules.

The usage of the testing framework requires Mininet, POX Controller, and any Linux operating system, although Ubuntu is recommended by the Mininet documentation. We used a pre-packaged Ubuntu and Mininet (version 2.2.2) virtual machine for this study. Aside from Mininet and POX, the tools used in the framework are: Tshark for the IP and MAC address extraction script, the Python module Scikit-learn for model training and validation, and hping3 for sending packets.

A command line interface tool was made to help facilitate invocation of specific framework functionality. This provides a clean abstraction layer from the intricate details and implementation of the framework's core modules. While out of the box it comes with four main functions, the CLI tool was designed to be extendable. Each CLI command has an accompanying command driver Python module. These drivers are then scanned by the CLI tool and put in hash table for quick access.

A global config file was provided for easy configuration of certain variables that the framework uses. For simplicity, the config file is in yaml format with a clear hierarchical structure for modifying variables. Available variable modifications include training data files directory, network generators, etc.

3.2.1 Model Training Module

A default model training module was provided as part of the framework. Users may use their own alternative Python module for training, given that he/she changes the config file to use that module instead of the default provided. Training modules should be placed in the ml_ids directory. We designed the provided training module to work for general IDS classifier training use cases and for rapid testing of different classifiers and datasets. As such, features to be extracted from the provided dataset are configurable using the global configuration file. Extracted features are then preprocessed through value normalization and one-hot encoding of categorical features. The preprocessed data is then fed into the model for training. The expected output of the module is a Pickle file containing the trained model, and cross validation test results using a portion of the training data as test data.

3.2.2 Model Validation Module

Given the necessity of validating trained models, we included a module that cross-validates a model using a different dataset but with the same features as specified in the configuration file used for training. This validation dataset is provided by the user, and is placed in the validation_data folder. It outputs the F1 score, recall, confusion matrix, and precision score of the validation procedure. This module is included to make sure that the trained model to be used in the IDS has not been overfitted, but the results here do not guarantee that the model will perform just as well when used in the real world. A user-defined model validation method may also be used by placing it in the ml_ids directory and changing the config file so that the CLI tool will use that module instead (Fig. 3).

3.2.3 Network Generator

A network generator was necessary to dynamically create a network topology in Mininet that would automatically scale according to any user specification. Separate packages were made for the internal and external parts of the network for modularity. The default network generators were made to recreate large and complex networks with ease. For this experiment, PCAP files were used to extract and replicate the network model of the PCAP.

MAC and IP addresses are extracted from PCAP files and are put in a hash table. MAC-IP pairs are then ordered by subnets with a prefix of 24. For the internal network, each unique MAC-IP pairs were given an associated host node in the network. Each host was then attached with an accompanying switch to act as the attachment

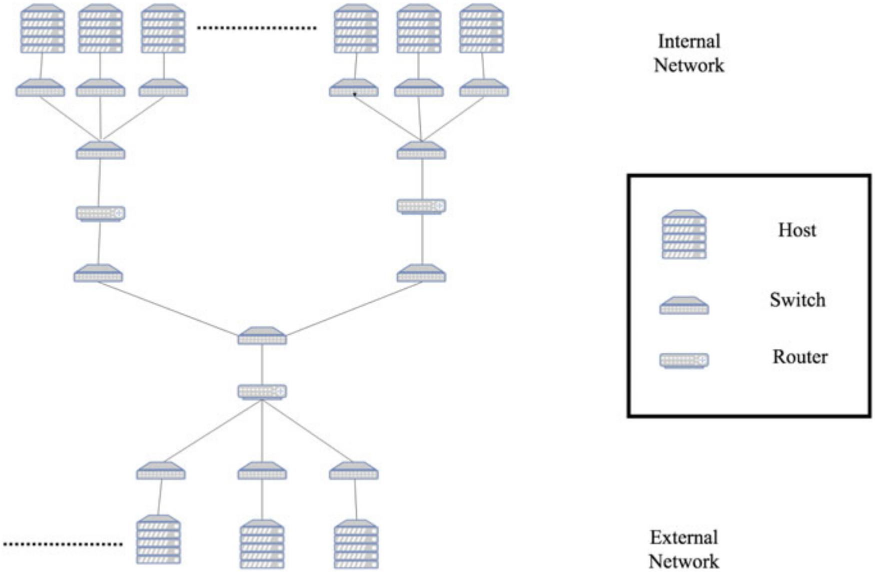


Fig. 3 Network setup diagram

for the IDS POX component. Each subnet is then given a router for connecting with hosts outside its subnet. The resulting network topology of the internal network generator should mimic the setup of any given dataset. Since most IDS testing would involve attacking target hosts, it was necessary for the internal network to have a 1:1 ratio of hosts and MAC-IP pairs.

For the external network, we opted to cut down the number of hosts based on the number of unique MACs. The external network module has a nearly identical generation process as the internal network module but with a few modifications for it to scale with large networks. After MAC-IP pair extraction, hosts are generated based on the number of unique MAC addresses. These hosts are given an arbitrary IP address in the same subnet with a prefix of 24. The extracted IP addresses with the same MAC addresses were then aliased on a host that contains their associated MAC address.

In our initial tests of a 1:1 host to MAC-IP pair setup, Mininet was unable to cope with the 24 000 hosts required by our dataset. As such, we opted to use the aliased IP setup to reduce the number of hosts from 24 000 to a manageable 8 hosts. This works for a PCAP replay experiment, in which it is not necessary to have a dedicated host for each MAC-IP pair. In our testing, it was sufficient to have the IP addresses aliased to a host. Like the other modules of the framework, both generators are swappable. Configuration changes for the default network generators were also integrated in its development.

3.3 Test Cases

Test cases are user defined Python modules located in the test_cases package. Test cases are expected to have a run method with at least five parameters. Each parameter corresponds to a collection of network component objects extracted after network generation. This allows the user to have complete access to all essential nodes in the network for facilitating tests of any type. As a sample, we developed a test module that executes a DDOS attack to a user specified set of hosts. This sample DDOS module was also used as the basis of benchmarking the framework itself.

The tests were run using the default distributed network topology included with the framework. Generated external hosts were then used as DDOS hosts. Each DDOS host was scripted to run *hping3—faster—c 10000000 targethost* to send 1000 packets per second with a hard limit of ten million packets. Background traffic was generated by running *hping3—c 6000 targethost*. This sends up to 6000 packets at a normal rate.

3.4 Global Configuration File

The global configuration file, which can be found in the config directory, is used to configure the different components or modules of the testing framework. The configurations for the training module, the validation module, the network generation module, and the CLI tool are set in this file. The folder name containing the datasets to be used for training and validating are set in the training-dataset-folder and validation-dataset-folder fields respectively. One must also set whether the data is in json or csv for both datasets in training-data-filetype and validation-data-filetype. For validation, one must set the classifier field to the model to be validated (models are located in the ids_models directory). For the training and validation modules, the feature names must be set for the source IP address, destination IP address, source port, destination port, protocol, packet count, and label. The network generation may be configured by setting internal and external network modules on the internal-network and external-network fields. Finally, the CLI can be configured so that the user may set the training module and validation module to be used by changing the training-module and validation-module fields respectively.

3.5 POX Components

The framework makes use of several POX components, one of which is the IDS to be tested. For this study, we made adjustments to the Guevara et al.'s IDS [2]. Instead of a threshold counter, we used the our trained classifier as the core of the detection system. This component is replaceable by the user. A user may provide his/her own IDS or modify the one provided, depending on his/her needs.

Another important component is the switch controller, used in the network. Each Mininet switch is connected to a controller, which implements switching in the SDN. Another component included in the framework is a metric logger. It is responsible for printing out the test results namely, F1 score, precision score, recall, and the confusion matrix, in a text file. To produce these results, it reads a text file that contains the IP addresses of the attack hosts, and checks whether these hosts were blocked by the IDS. It also logs all blocked IP addresses in the console and compare these logged addresses to the list of attack hosts.

3.6 Procedure

The general flow for this test framework uses POX and Mininet in separate terminals. POX is first started up together with the custom components as command line arguments to boot up the controller. Using the CLI tool in another separate terminal, the testing network topology may be set up. The user may also optionally pass the name or pointer to a test case as an argument when generating the network environment to run those test cases. The CLI tool is run by entering the following in the terminal: [sudo] python tool.py [command] [other_args]. The available CLI commands are "train," "validate," and "createnetwork." The custom Python script will generate the required hosts and switches for both the internal and external networks, it will then run the included tests. After the tests have been run, the results are outputted in text files which can be found in the results directory.

The user may also use the CLI tool to train and validate models to be used in their IDS. The user must have configured the training and validation modules in the configuration file, and have placed training and validation datasets in their respective directories. After setting the proper configurations, the user may run the training and validation modules using the "train" and "validate" commands of the CLI tool respectively. Models produced by the training module can be found in the ml_ids/ids_models directory. The user may train and validate as many models as they want using the CLI tool.

4 Results

See Table 1.

Initial tests were conducted to verify the validity of the testing framework. We used an Adaboost classifier, trained using the ISCX 2012 Dataset, as the core of the IDS for all scenarios. The dataset features we used mimics the typical extractable data found from inspecting packets, namely: source ip and port, destination ip and port, protocol and counters for total packets from the same source. Each entry of

Table 1 Test plan

Test objective	Setup	Expected Result	Result
Normal traffic only	20 background traffic hosts sending normal traffic to target hosts	The IDS would not find any anomalous packets	No hosts were tagged as anomalous Precision: 1 Recall: 1 F1 score: 1
DDOS attack with normal traffic	7 DDOS hosts flooding 3 targets 20 background traffic hosts sending normal traffic to target hosts	The IDS would detect attacks amidst normal traffic	IDS did not detect 2 out of 7 attack hosts Precision: 1 Recall: 0.71 F1 score: 0.83

Table 2 IDS classifier training confusion matrix

	Negative	Positive
Negative	43 765 453	19 236
Positive	55 963	3 775 975

the dataset is then labeled with a tag to identify if the packet is anomalous or not. Validating the trained classifier yielded a precision of 99.49%, a recall of 87.09% and an F1 score of 92.88% (Table 2).

The first test case involves testing the IDS' ability to ignore normal traffic. No hosts were tagged as anomalous, resulting in a precision of 1, a recall of 1 and an f1 score of 1. This was used to test if the classifier overfit with the training data or not.

The second test case involves testing the IDS' ability to detect attacks with normal traffic running in the background. Combining these two was necessary to perform a more realistic test scenario. This specific test resulted in a precision of 1, a recall of 0.71 and an f1 score of 0.83. It missed two of the seven attack hosts and classifying all background traffic hosts as normal. The framework was able to capture the results of all the test scenarios correctly. All test scenarios used the metric logger as a way to track blocked hosts and to compute for precision and recall as well as output the resulting confusion matrix.

5 Conclusion

5.1 Summary

In this study, we set out to create a framework for testing machine learning IDSes in SDN. We also wanted to give the framework's users a set of tools that will help them in creating their own machine learning IDSes. These objectives were achieved by: creating a Python script that automatically generates a network environment given MAC-IP pairs; creating a script that allows ripping MAC-IP pairs from PCAP files;

providing a POX component that logs detected attacks and compares those to the list of true attacks to provide results of the IDS test; allowing the user to provide their own test cases and to automatically run those tests after generating the network topology; making the the framework highly configurable; creating Python scripts for training and validating models to be used for their IDSes.

We modified an existing statistical IDS and turned it into a machine learning IDS using our own model. The model was trained using the ISCX 2012 Intrusion Detection Dataset, then validated using the Coburg Intrusion Detection Dataset.

This study resulted in a configurable and extendable testing framework for Machine Learning IDSes.

5.2 Recommendations

For future researches, we recommend that the framework include more metrics that measure the performance and effectiveness of IDSes. In its current state, the framework only measures the accuracy of IDSes in detecting attacks. Another recommendation is to make the modules for training models easier to configure, so that switching between different classifiers would be much easier.

References

1. Open Networking Foundation (2013) Software Defined Networking (SDN) Definition. https://www.opennetworking.org/sdn-resources/sdn-definition
2. Guevara AG, Domingo MA, Yu, W (2014) Enhancing intrusion detection and prevention systems using software defined networking technology in a distributed topology. Undergraduate thesis, Ateneo de Manila University
3. Pena JG, Yu W (2014) Development of a distributed firewall using software defined networks (SDN) technology. Undergraduate thesis, Ateneo de Manila University
4. Mell PM, Lippmann R, Chung TH, Haines J, Zissman M (2003) An overview of issues In: Testing intrusion detection systems
5. Dongre S, Wankhade K (2012) Intrusion detection system using new ensemble boosting approach. Int J Model Optim 2(2012):488–492
6. Kumar G (2014) Evaluation metrics for intrusion detection systems—a study. Int J Comput Sci Mob Appl 2014:11–17
7. Munaiah N, Meneely A, Wilson R, Short B (2016) Are intrusion detection studies evaluated consistently? A systematic literature review. RIT Sch Works 2016:1–18

Part VI
Data Mining and Artificial Intelligence

Ballet Pose Recognition: A Bag-of-Words Support Vector Machine Model for the Dance Training Environment

Margaux Fourie and Dustin van der Haar

Abstract Serious dance students are always looking for ways in which they can improve their technique by practising alone at home or a studio by using a mirror for feedback. The problem these students face is that for many ballet postures it is difficult to analyze one's own faults. By not having guidance regarding proper positional alignment, dancers risk developing injuries and bad habits. The proposed solution is a system which recognizes the ballet position being performed by a dancer. After recognition, this research aims to work towards providing the necessary correction as feedback. The results for recognition in the system, using a Bag-of-Words approach to a Support Vector Machine classifier, showed an accuracy of 59.6%. Multiple implementations are produced and assessed in this paper. It is clearly found that the approach is feasible, however, work for improving the accuracy is required. Recommendations to improve effective pose recognition for future work are therefore discussed.

Keywords Posture recognition · Computer vision · Ballet training · SVM

1 Introduction

In an effort towards continuous improvement, many dancers practise by themselves, but during these sessions, they often miss having the feedback of a dance teacher. The problem this paper aims to address is that training without any guided assistance, dancers are less likely to spot their own improper alignment, which, if not corrected, leads to the development of bad placement habits and a higher risk of getting injured. In his book on Conditioning for Dance, Franklin clearly indicates that dancers often

M. Fourie · D. van der Haar (✉)
Academy of Computer Science and Software Engineering, University of Johannesburg,
Cnr University Road and Kingsway Avenue, APK Campus, Johannesburg 2006, South Africa
e-mail: dvanderhaar@uj.ac.za

M. Fourie
e-mail: margauxf@uj.ac.za

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_33

revert to dance positions that feel comfortable, but are misaligned and inefficient for the required movement [1]. He also points out that once dancers can break any hardwired habits of misalignment, they are able to reach a new level of mastery while reducing the risk of potential injury [1]. The proposed project will approach this problem with the intention of enhancing the training lives of ballet dancers, by starting with pose recognition. With an additional tool for proper training, dancers may further develop their technical abilities which are crucial in today's demanding ballet world [2]. The problem of unguided training is also relevant with regard to dance teachers who are responsible for frequently exploring new ways to pass on training ideas in order to produce well-trained dancers [3]. The proposed system may, therefore, eventually be used as an additional form of feedback.

Currently, the research that exists in the field of pose recognition and correction for ballet training is still limited, which is why it is relevant to explore various methods to approach this problem [4]. This paper conducts a literature review in Sect. 2. Section 2.1 starts with the problem background. Section 2.2 investigates relevant fields and research that is currently being explored for ballet posture recognition. The experiment setup is described in Sect. 3. The proposed model and it's implemented algorithms are unpacked in Sect. 4. Section 5 contains a discussion of the results and potential ways to improve the model. Finally, Sect. 6 concludes the paper with the vision of potential future work in the problem space of ballet pose recognition.

2 Literature Review

2.1 Problem Background

Classical ballet originated in the 16th century. Over the years since then, training methods for this athletic art form have remained mainly the same. The knowledge base for ballet training has always been dependent on teachers passing on their expertise to future generations [5]. The traditional environment in which ballet training occurs can generally be described as a studio classroom setting with approximately 8–10 dance students and a single dance teacher. The ballet teacher can however not keep his/her eyes on everyone in the classroom at the same time. For this reason, private one-on-one coaching is often essential and of great benefit to the serious ballet student [6]. The demand for guided individual training is continuously increasing, which hints at the use of technology in the dance environment to address this requirement. In order to properly address the recognition and correction of ballet poses, there is value in understanding the concept of ballet technique. It can be described as the fundamental principles for body lines, form and movement in the ballet-style [7]. Ballet technique, therefore, deals with various aspects of the body such as alignment, turnout, posture, as well as stretched legs and feet [7]. A closer look will be taken at relevant research in the subsection that follows.

2.2 Related Relevant Fields

How well a ballet dancer executes their technique has always been dependent on the correct placement of their ballet postures. Currently, there are multiple ballet posture training systems being explored that make use of different technological methods which include, wearable technologies as well as systems using camera hardware, such as the Microsoft Kinect [8].

Wearable Technology
An interesting area of modern technology that is being explored for ballet training is wearable devices or garments. One such approach, involves a teacher wearing a full-body garment which is geared towards helping inexperienced adult ballet beginners understand proper technique better [9]. The garment lights up the essential limbs being used by a teacher which enables beginners to follow along and focus on the most important aspects of movements [9]. Some disadvantages of this system, however, include the high cost of constructing such a garment as well as the restriction of movement [9]. The keypoint-based instruction method of this wearable technology approach is an effective way to coach various levels of ballet.

Computer Vision and Choreography
The field of computer vision can be applied to the scenario of ballet choreography which involves the creation of dance sequences, built up from ballet technique steps and floor patterns within the dance space [10]. Dancs et al. explored the choreographic side of ballet with a project aimed at automatically recognizing and recording a choreographer's movements. The study made use of Microsoft's Kinect in order to detect the joints of the body. It also used various classification algorithms such as Nearest Neighbor (NN) methods and Support Vector Machine (SVM) methods. Despite limitations such as the Kinect sensor not detecting particular ballet stances, the approach yielded promising results with the main algorithms generating results over 90% for accuracy [10].

Posture Recognition in Ballet
A research field which largely forms the basis for the proposed project is referred to as posture recognition. One early approach by Saha et al. designed a fuzzy algorithm that matched and recognized ballet postures. The work involved various preprocessing methods which reduced the dancer image to a stick figure representation and thereafter the proposed algorithm was used to perform recognition [11]. This initial project, produced an accuracy rate of 82.35% for the recognition of stances.

3 Experiment Setup

The proposed project deals with the recognition of the specific ballet pose called Retiré or Passé. In this paper, we will refer to it as the Retiré position. Therefore, this implementation is a binary classification problem which identifies whether a dancer is in the position or not. A Retiré is the position in which most "spin" movements (called Pirouettes) occur in ballet. The pose can be described as the dancer standing on one leg, with the other leg pulled up to the knee of the supporting leg. By perfecting this pose, dancers will be able to perform turns with more stability [12].

The setup of the proposed project generally consists of having a dancer captured in an image by using a webcam or an offline image from a dataset. For pre-processing the first step is to gray-scale images, followed by histogram equalization and body detection using the histogram of oriented gradients approach. For the feature extraction step, SURF features were extracted from the pre-processed images. The last part of the experiment setup involved training and testing a Bag-of-Words based Support Vector Machine.

Currently, the data for the proposed system is sampled from offline datasets containing images of dancers in various positions before, during and after they perform a Pirouette movement. This implementation originally made use of a limited dataset sourced from the action dataset of the University of Central Florida [13]. This dataset contained low-resolution images which produced the initial results. Thereafter, the larger dataset, with higher resolutions, was compiled from video frames with the aim to gather more metrics on the model's results. The environment in which this system would be used is essentially anywhere with sufficient space for a dancer to practise poses. The constraints that have to be in place for this particular system are as follows: No mirrors should be in the captured image in order to eliminate reflection; minimal objects should be in the background of the captured image and, furthermore, the lighting should not be too poor or too extreme. The above constraints ensure the optimal conditions to capture and recognize the Retiré position.

4 Proposed Model

The pipeline that was followed for this model generally consisted of the high-level pattern-recognition steps namely: pre-processing, feature extraction and classification. The pre-processing steps for this model included the gray-scaling of images, followed by histogram equalization. Next, a histogram of oriented gradients approach was used for detection of the dancer's body. For feature extraction the Speeded-Up Robust Features (SURF) method was used. The classification step used a Bag-of-Words approach together with a Support Vector Machine classifier. In the section below, an inspection of the algorithms involved in each step of the project will follow (Fig. 1).

4.1 Pre-processing

Gray-scaling of the images is the first step in the pipeline of the proposed model in order to save computation time and convert the images to a simpler color space.

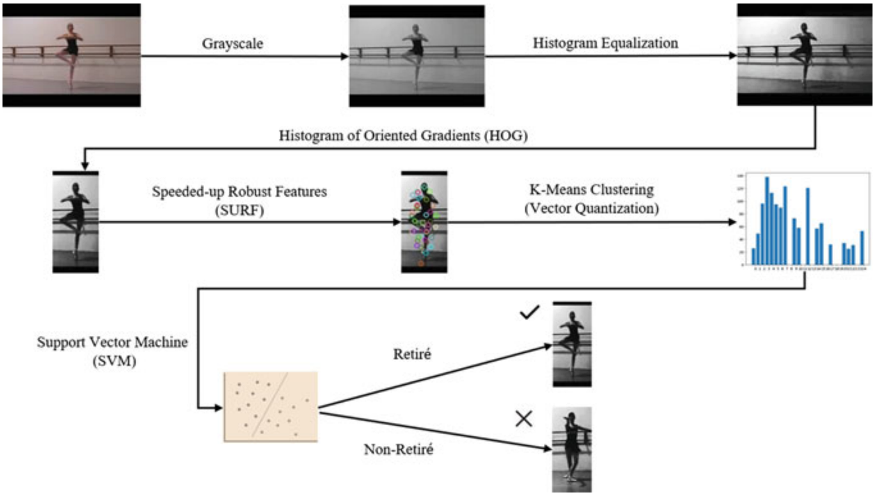

Grayscale
Histogram Equalization
Histogram of Oriented Gradients (HOG)
Speeded-up Robust Features (SURF)
K-Means Clustering (Vector Quantization)
Support Vector Machine (SVM)
Retiré
Non-Retiré

Fig. 1 Visual representation of the methods implemented for the proposed model

Histogram Equalization is used as the next pre-processing technique to bring out the contrast in the images more clearly. Bradski et al. describe Histogram Equalization as involving the mapping from an original intensity distribution to a new, more evenly spread out, intensity distribution [14]. The result is an image with stronger enhanced contrasts [15], which highlights the outlines of dancers' bodies more clearly. The next algorithm, called Histogram of Oriented Gradients (HOG), detects the body of the dancer in an image by using the differences in edge intensities. The distribution of the directions of gradients forms a histogram of features that is used to ultimately detect where the body is in the captured image [16]. After HOG is completed, the images are re-sized to a uniform resolution to ensure that fair and accurate feature extraction can take place in the next step of the pipeline.

4.2 *Feature Extraction*

For the proposed Model's pipeline, it was decided to make use of the Speeded Up Robust Features (SURF) Algorithm. This method includes all the advantages of the Scale Invariant Feature Transform (SIFT) method; however, it is faster which saves computation time. The benefits of using SIFT and SURF for feature extraction include that they are scale-invariant as well as rotation invariant [14]. Furthermore, these methods reduce the probability of poor feature extraction as a result of clutter, noise or occlusion [17]. After feature extraction, it is necessary to provide a classifier with these features for training and testing which is discussed in the section below.

4.3 Classification

A Bag-of-Words approach with a Support Vector Machine (SVM) was used for classification in the proposed system. In this classification approach, the feature vectors are quantized by using K-means clustering. It was decided that 25 clusters best suited the model for the initial dataset. Each of the features detected in the images for training was then placed under one of these clusters, ultimately creating a histogram of features to be provided to the SVM classifier. Before the histogram of features is used to train the SVM classifier, the histogram is standardized to normalize the distribution of features. This prevents the classifier from becoming biased as a result of steep variances. Bradski et al. point out that when one's dataset is limited, an SVM algorithm, is one of the best approaches to use for classification [14]. It is therefore because of a limited dataset that this model proposes the use of SVM as an appropriate classifier. Once the training of the SVM classifier is completed, images from the test set are processed and its features are used to test if the SVM accurately predicts whether a dancer image is in the Retiré position or not. The results and metrics gathered from the testing phase of the Support Vector Machine are discussed in the next section.

5 Results

The results of this proposed system indicate how well the model is able to recognize the Retiré position when test images are provided. Generally, the collected results hint at the fact that there is room for improvement of the model. An important aspect in terms of the model's performance was the data provided to the classifier, which is why two different datasets were tried and tested. The training and testing split that was used on the datasets was 40% for training and 60% for testing. The Bag-of-Words Support Vector Machine (BOW-SVM) classifier which was trained on the initial smaller dataset achieved an accuracy score of 59.6%. The misclassifications (where the model's predictions were incorrect) were largely due to dealing with limited data and low-resolution images. Considering this, the model fares quite well and is feasible since it performs better than a random classifier. Furthermore, it can be observed that the model's specificity for detecting negative instances was high. The recall rate, however, indicated a low sensitivity to detecting positive instances. The results of the Bag-of-Words approach SVM on the smaller dataset are summarized in Table 1 (Table 2 and Fig. 2).

The Equal Error Rate (EER) of the system is the value where the false positive rate and the false negative rate is equal. It is desirable to have the EER as low as possible to ensure a higher accuracy. The Equal Error Rate for the initial dataset is high due to a limited amount of data and therefore has the potential to improve.

In order to investigate more metrics on the BOW-SVM classifier, another classifier was trained on a larger dataset consisting of slightly higher resolution images. The

Table 1 Summary of results on smaller data set

Metric	Result in %
Accuracy	59.6
Misclassification rate	40.4
Recall rate	4.2
Specificity	100
False positive rate	0
Precision	100

Table 2 Equal error rate for smaller data set

Recognition model	EER %
BOW SVM	45.5

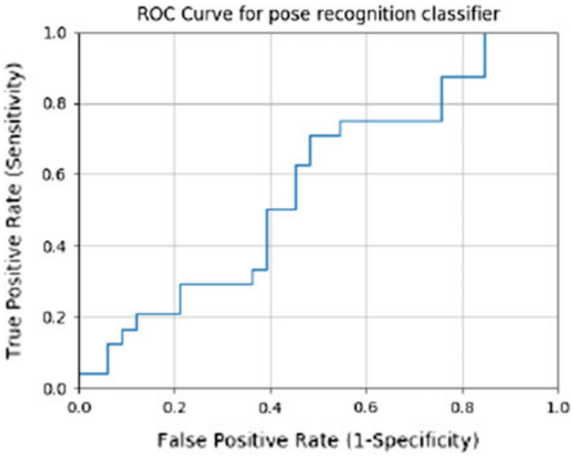


Fig. 2 ROC curve for smaller data set

initial accuracy for this classifier was 46.3%, which hinted at improvements required for future work, such as tuning the pre-processing parameters to better suit the larger dataset of images. The recall rate indicated a lack of sensitivity to positive instances; however, the specificity of the model reached 97.6%, indicating that detection of the non-Retiré instances is good. The results of the classifier trained on the larger dataset are summarized in Table 3. The Equal Error Rate for the larger dataset is lower due to the classifier being trained on more data. The value of 28% is better when compared with the initial dataset's EER (Fig. 3 and Table 4).

A similar system that also had an SVM approach by Saha and Konar in 2015 reached an overall accuracy result of about 79% for recognizing particular poses [19]. Their system, however, did not recognize for the Retiré position. When one compares the performance of the two classifiers discussed above, it is clear that the first SVM model had a higher accuracy level. A reason for this is that the pre-

Fig. 3 ROC curve for larger data set

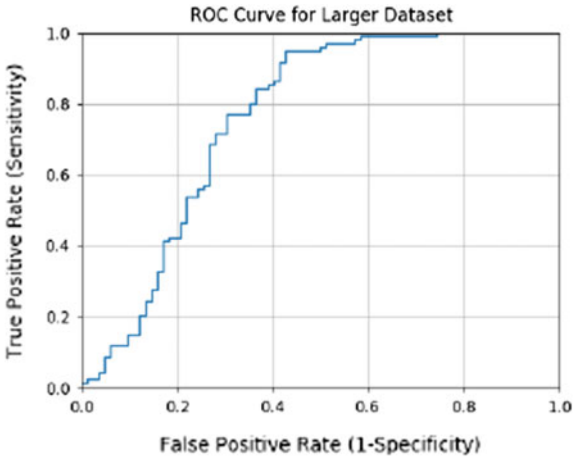


Table 3 Summary of results on larger data set

Metric	Result in %
Accuracy	46.3
Misclassification rate	53.7
Recall rate	2.1
Specificity	97.6
False positive rate	2.4
Precision	50

Table 4 Equal error rate for larger data set

Recognition model	EER%
BOW SVM	28

processing techniques used in the pipeline were well suited for the smaller dataset's images. Once sufficient pre-processing and feature extraction can be performed on the new dataset, the second model's accuracy should improve significantly. The resolution of images plays another important role in producing accurate classifiers, which is why an attempt to sample more, high-resolution data is essential for future work of improving the above-mentioned models.

6 Conclusion and Future Work

The proposed system discussed in this paper has been an interesting investigation into a unique approach towards pose recognition for the particular Retiré ballet pose.

The one implemented SVM classifier reached an accuracy level above that of a random classifier, whilst using a small dataset containing low-resolution images. Future work will, therefore, include sampling more, higher resolution data in order to improve the classifier's performance. Another future improvement that can be made on the current system is to fine-tune the pre-processing methods in order to provide the classifier with more SURF features. The discussed pipeline, therefore, aimed to approach the problem of recognizing a ballet pose in a unique way by using various image-processing methods together-with classification algorithms. The work proposed in this paper strives towards a pose recognition system that can, in future, effectively assist ballet dancers in their dance training.

References

1. Franklin E (2004) Conditioning for dance: training for peak performance in all dance forms. Human Kinetics
2. Speck S, Cisneros E (2003) Ballet for dummies
3. Kassing G, Jay DM (1998) Teaching beginning ballet technique. Human Kinetics
4. Banerjee A, Saha S, Basu S, Konar A, Janarthanan R (2014) A novel approach to posture recognition of ballet dance. In: 2014 IEEE international conference on electronics, computing and communication technologies (CONECCT)
5. Trajkova M, Cafaro F (2016) E-ballet: designing for remote ballet learning. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing adjunct—UbiComp '16, pp 213–216
6. Spirit D (2014) Working one-on-one: what to expect from private lessons
7. Dancing RA (1997) The foundations of classical ballet technique. Royal Academy of Dancing
8. Hong GS, Park SW, Park SH, Nasridinov A, Park YH (2016) A ballet posture education using it techniques. In: Proceedings of the sixth international conference on emerging databases technologies, applications, and theory—EDB'16, pp 114–116
9. Gupta M, Hallam J, Keen E, Lee C, McKenna A (2014) Ballet hero: building a garment for memetic embodiment in dance learning. In: Proceedings of the 2014 ACM international symposium on wearable computers adjunct program—ISWC '14 Adjunct, pp 49–54
10. Dancs J, Sivalingam R, Somasundaram G, Morellas V, Papanikolopoulos N (2013) Recognition of ballet micro-movements for use in choreography. In: 2013 IEEE/RSJ international conference on intelligent robots and systems, pp 1162–1167
11. Saha S, Banerjee A, Basu S, Konar A, Nagar AK (2013) Fuzzy image matching for posture recognition in ballet dance. In: 2013 IEEE international conference on fuzzy systems (FUZZ-IEEE)
12. BalletHub (2017) Passe and retiré basics
13. CIL U (2011) Ucf-cil action dataset @ computational imaging lab-university of central florida
14. Bradski G, Kaehler A (2008) Learning OpenCV: computer vision with the OpenCV library. O'Reilly Media, Inc
15. Gonzalez RC, Wood RE (2008) Digital image processing, 3rd edn. Prentice-Hall
16. Bhangale K, Shekokar R (2014) Human body detection in static images using hog & piecewise linear svm. Int J Innov Res Dev **0**(0)
17. Nixon MS, Aguado AS (2008) Feature extraction & image processing for computer vision. Academic Press
18. Muneesawang P, Khan NM, Kyan M, Elder RB, Dong N, Sun G, Li H, Zhong L, Guan L (2015) A machine intelligence approach to virtual ballet training. IEEE MultiMed 22(4):80–92
19. Saha S, Konar A (2015) Topomorphological approach to automatic posture recognition in ballet dance. IET Image Process 9(11):1002–1011

Pronunciation Detection for Foreign Language Learning Using MFCC and SVM

Jihyun Byun and Dustin van der Haar

Abstract As technology improves, people around the world are given more effective tools to communicate with each other. This has caused a sensation of secondary language learning. Many countries have now included this as an obligatory component of their education systems. However, the lack of appointing right professionals has led to misleading the practicing the pronunciation of the new language, because students often follow the pronunciation that non-native teachers have. This paper aims to provide a model that has a potential to help learners with increasing the recipient for understanding the speaker. The model records the learner's English pronunciation of a given context, analyses it and provides feedback on the screen. The system has shown an accuracy of 98.3%. Throughout the research we have discovered that several factors such as the learner's predefined accent from his mother-tongue language, the noise level of an environment where the learner uses the system as well as different types of English accents interfere with providing accurate feedback to the learner.

Keywords Signal processing · Pronunciation · Biometrics · Education

1 Introduction

Speech is one of the natural parts of being a human. There is a rarely a day without speaking; people use this verbal expression to share their thoughts and emotions. Moreover, speaking is a popular method for building social interaction. Therefore, it is not surprising to state that speaking plays a vital role in human society.

J. Byun · D. van der Haar (✉)
Academy of Computer Science and, Software Engineering, University of Johannesburg, Cnr University Road and Kingsway Avenue, APK Campus, Johannesburg 2006, South Africa
e-mail: dvanderhaar@uj.ac.za

J. Byun
e-mail: ashburncrash@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_34

With the emergence of globalization, the human interaction has extended from the local area to across the countries. Rapid improvement of technology has enabled the effective communication between foreigners. In order to keep up with this global trend, many countries have adopted a module for learning second languages to their educational systems. This is exciting, however, it is worth mentioning that many second language (L2) learners struggle with speaking while showing great strength in other sectors such as reading, listening and writing.

During our research, we have found several problems with learning a foreign language with non-native speakers. First, the language is often not taught by a native speaker, unless the learners reside in a home country of the language that they wish to obtain. Instead, the teachers who themselves learned it as a second language is usually appointed. This causes huge variation and incorrectness in pronunciation, which is a prominent part of learning a foreign language. Secondly, a learner tends to feel intimidated by his non-perfect pronunciation and possible criticism from a teacher or a native speaker. This may leave the learner in a discouraging situation, depriving his aspiration. In this paper, we introduce a model that assists the learners such that they can successfully acquire accurate pronunciation of the language.

The paper is designed as follows: Sect. 2 discusses the problem background, while Sect. 3 explores similar systems of the topic. Section 4 contains the details of a proposed solution for the problem and its results are discussed in Sect. 5. Finally, Sect. 6 concludes the article.

2 Problem Background

The concept of education—passing knowledge from older generation to the newer generation—has existed for a long time. For many years, conventional education system consisted of a teacher and one or more students. The lessons conducted were usually led by the teacher in a face-to-face classroom environment [1].

For language learning, many changes have taken a place. In the beginning, non-native teachers were often assigned to teach a second language course. Grammar and reading comprehension were main focuses during this phase. Then, non-native teachers with some experience with the corresponding languages were appointed for the tutoring. The speaking section is then introduced, but problems start to emerge. Often the teachers would have wrong pronunciation, and this was taught directly to the students. Moreover, accents or dialects were restricted to the teacher's choice.

It was not so long ago when educators realized the importance of speaking in learning the second language. In many institutions, non-native teachers were replaced with native teachers. Acquiring correct pronunciation became a staple goal. Students have trained the basic elements such as intonation and stress on words. Free-topic chat, role-plays were pioneered to familiarize students with the language, as well as to instill confidence in speaking. Unfortunately, the learning only lasted within the institution—in other words, there was no appropriate tools or environment for students to practice L2.

In traditional language learning, students heavily depended on teacher's feedback for pronunciation. Methods such as listen-and-follow teacher's pronunciation or minimal pair drills were used [1]. Often, students were recommended to record their pronunciation to detect a problem in their pronunciation. However, according to an experiment conducted by Foote [2], this did not turn out to be efficient to improve one's skill. A different approach was required to help the students.

Language learning requires continuous effort from students to achieve fluency in a language. Several studies such carried by Furtak, Kunter and Dickinson have shown that autonomous learners have high motivation for accomplishing learning objectives [1]. By taking accountability for their own learning, a learner can treat learning as an experiment, which allows them to explore which way is suitable to yield the best result. In this process, receiving feedback thus has a significant impact on the learner's performance—and technology can help.

Kruk's experiment [3] showed the effectiveness of using online resources for language pronunciation. Among three test groups, a group that was instructed to utilize autonomy in language pronunciation has an outstanding result over other groups, which were traditional teaching based and controlled without any specific instruction. However, Kruk warns that autonomous attitudes should be gradually introduced as some students may not deal with sudden changes very well. He also suggests autonomy in teachers [3].

Automatic Speech Recognition (ASR) has been a popular technique for various applications, such as voice and speaker recognition [4]. In voice recognition, the acoustic input from a user and provides feedback on the words spoken as text. Common applications in current society include Siri, Google Assistant and many other language programs such as Rosetta Stone have implemented ASR as part of language learning process. ASR has been criticized for its inaccuracy, however, it has been improving. The limitations are mainly due to careless design or unfamiliarity with the system and ASR can be used to achieve correct error detection and giving a feedback [5]. In Hincks' paper [6], ASR is useful for teaching beginner-level learners, but it has not yet reached for delicacy and sophistication of human language speech.

3 Related Works

SLAP

Second Language Acquisition Pronunciation, or SLAP, is a system that was suggested by Gu and Harris [7]. It simulates an interaction between a learner and a native speaker for learning English as a second language. The system automatically detects mispronunciation and repeats it back to the learner, emphasizing on the part that has been incorrectly pronounced. SLAP aims for robustness in detection, especially with complex words. SLAP system utilizes Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) algorithm. The system takes 256 samples to create MFCC feature matrix (N × M), where N is 13 feature coefficients and M is the number of windows per utterance. Dynamic Time Warping (DTW) algorithm estimates the time difference between two pronunciations, the learner's input and stored native speaker's pronunciation before they are compared. Additionally, local constraints, global constraints, and slope weighting are applied to control for a search on the optimal path.

SLAP spots mispronunciation by comparing the differences between maximum local distance value per windows on the optimal path. To emphasize the incorrectly pronounced part, the maximum value is obtained from local distance on optimal path. The value is used to find the most unmatched part of the utterances. Then, Linear tapering is used to smooth out the modified pronunciation. SLAP has shown the good performance in detecting mispronunciation. However, the system has lack of participants during testing and poor user diversity. In the next section, we will present a pronunciation assisting model that factors in these aspects.

There are other systems that aim for improving non-native English speakers' pronunciation. Such systems include CMUSphinx, ELSA, Duolingo, and FLUENCY. However, since they are commercialized products, their pipelines and corresponding algorithms are not revealed to the public.

4 Proposed Model

We propose a system that detects and analyses the pronunciation of the learners who are in the process of learning a foreign language to improve speaking skills. A learner will pronounce a given word and it would be compared to native speaker's pronunciation, that is stored in a database. Depending on a result of the analysis, the system will return a feedback which may require the learner to repeat the pronunciation.

This digital language learning assistant will be available to anyone, regardless of time and space. English was chosen as a test language for this purpose, however, the system aims to provide a general algorithm that can be applied not only a chosen test

language but also other foreign languages. The brief overview of the system pipelines is depicted in Fig. 1. The rest of the section will discuss each phase of the system and details of them.

The proposed system is built with Python programming language, with the help of Librosa, a python package for audio analysis. Other libraries such as scipy and numpy were chosen for mathematical computation, as well as matplotlib for visualizing the process, and sckit-learn for a Support Vector Machine with linear kernel.

Capturing User Input

A learner is instructed to pronounce a given context through a microphone attached to a desktop device. This raw input will then be conformed to a format we have predefined. We concluded to follow a standard format for human speech, which includes 16 kHz sampling frequency, 16-bit resolution, mono-channel and in an uncompressed format.

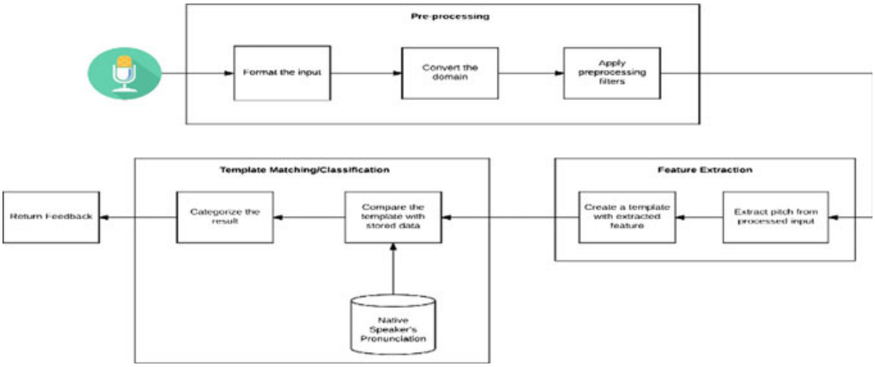

Pre-processing
Format the input
Convert the domain
Apply preprocessing filters
Template Matching/Classification
Feature Extraction
Return Feedback
Categorize the result
Compare the template with stored data
Create a template with extracted feature
Extract pitch from processed input
Native Speaker's Pronunciation

Fig. 1 The overview of proposed model

In the next section, we will discuss preprocessing techniques that are performed on the raw input. These techniques are implemented prior to feature extraction phase, that contributes towards the competent performance of the model.

Preprocessing

In this phase, front-end processing is applied to an audio sample that is captured by a user. The process includes pre-emphasis, multiplying by a windowing function and executing fast Fourier transform (FFT). Preprocessing is done to improve the quality of the raw input by filtering unnecessary noises so that it is ready for the feature extraction phase.

Segmentation and Windowing

First, we multiply a windowing function to the signal so that it can be prepared for Fourier Transform. To do so, we segment the samples into a number of frames. We have chosen 25 ms to be the length of a frame and 10 ms as the length of frame step. Such values are chosen as recommended frame length is 10–30 ms. A smaller value would have too little information to extract from the raw data, while the higher value would yield too much information and noises to extract. Frame step is an overlapping part between frames. Since our sampling rate is 16000, the frame length and frame step will be 400 and 160 samples, respectively. This indicates that first frame will start at 0, the next sample will start at 160th sample, and so on.

After segmentation, a window function will be applied to each frame. Windowing reduces spectral leakage, which is caused by discontinuities between frames of the segmented signal. Common options are Hanning or Hamming windowing. We have chosen the Hamming windowing, which is half cosine function to smooth out edges of each frame.

Fourier Transform

We apply Fourier Transform to convert time domain of our signal to frequency domain. Frequency domain provides a clear view of what the signal contains. Instead

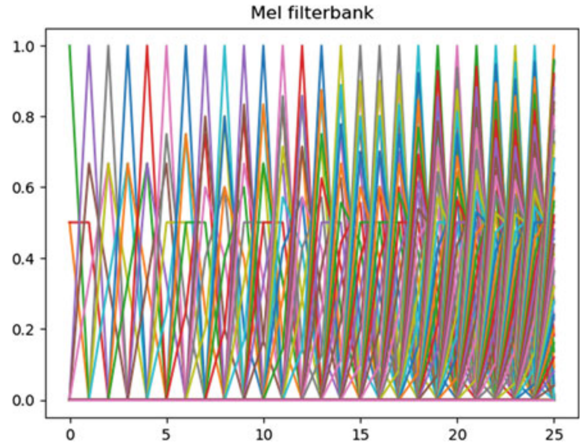


Fig. 2 Computed 26 Mel Filterbanks, where the x-axis is the filterbank and the y-axis the spectrum

of traditional Fourier Transform algorithm, Discrete Fourier Transform (DFT), we apply Fast Fourier Transform (FFT), which accelerates the process of frequency domain conversion. For the number of FFT or FFT bin numbers, we have chosen a default value, which is 512. FFT size is recommended to be bigger than the frame size.

Feature Extraction

There are numerous types of features can be extracted from a speech segment, such as pitch and energy. Before the introduction of Mel-Frequency Cepstrum Coefficients (MFCCs), Linear Prediction and Cepstral Coefficients were popular algorithms for speaker processing [8]. We have selected MFCC as a feature to be extracted. MFCC features are short-term spectral features (Fig. 2).

First, we will compute the Mel-spaced filterbank that contains 26 triangular filters. This filterbank is created with the lowest frequency of 300 Hz and highest frequency of 8000 Hz, which is half of our sampling rate. Each triangular filter is then multiplied to power spectrum that is produced as a result of FFT. This will yield 26 filterbank energies.

The logarithm form of these energies is calculated, and Discrete Cosine Transform (DCT) is applied. The reason for taking logarithm form is to alter the energies so that it matches closer to human hearing perception. Moreover, the logarithm form produces the channel normalization technique through cepstral mean subtraction [8]. DCT removes correlations between the logarithm of the energies so that the features can be used by a classification model efficiently. The outcome will be 26 MFCCs. However, we will be only keeping cepstral coefficients 2nd to 13th as high cepstral coefficients contain fast changes, which decreases the performance of speech recognition system [8]. Furthermore, the first index of the coefficient is discarded since it contains the information on the sum of the energies, which is not relevant in the system. Figure 3b depicts the resultant MFCC that we have computed.

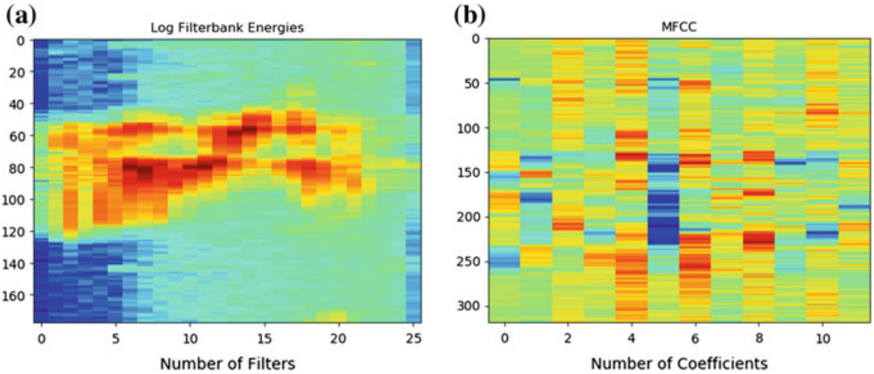


Fig. 3 The process of feature extraction. **a** Logarithmic computation on initial filterbank energies. **b** 12 MFCCs

Classification

In digital signal processing, Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) have been the dominant choices for many years. In this paper, we have selected the Support Vector Machine with linear kernel as a classifier of the proposed system. The classifier will be trained with native speakers' English pronunciation of the words from Wikimedia Commons dataset and non-native speakers' English words pronunciation from Berkeley Restaurant Project dataset. We have selectively reconstructed Berkeley Restaurant Project dataset to suit our needs. The combined dataset will be shuffled and be split in the ratio of 6:4, training and testing dataset size respectively. After the training, the classifier will be further assessed with the test dataset, which is a custom collection of both native and non-native English speakers' pronunciation.

5 Result

Initially, we started training our system with half of each dataset. The model yielded 98.3% accuracy. We added a new dataset called Mocha TMINT, a British speech corpus, to test whether the system would be able to perform differently with English accents. The resultant accuracy was 88%. Furthermore, the system yielded an equal error rate of 0.077%. The Receiver Operating Characteristic curve of the performance of the classifier (trained with 6000 audio files) is shown in Fig. 4 below.

Moreover, we have found that factors that impact accuracy include speaking in low volume, distance from the microphone or the student murmuring.

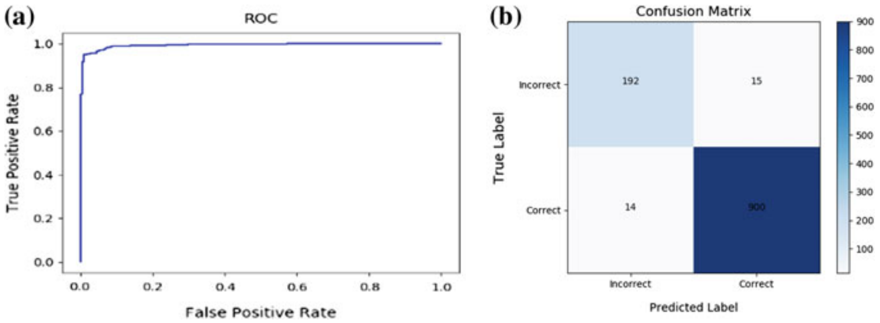


Fig. 4 **a** Receiver operating characteristic curve of our classifier. **b** Confusion matrix of our classifier

6 Conclusion

Digital signal processing explores one of the most vital functions in human beings, speaking. Through our research, we were able to learn about the complexity of human speech and various algorithms for analyzing it. MFCC has allowed us to extract features that are close to human hearing, which is helpful for training a machine that does not have prior knowledge of how a human perceives the sound. Our system distinguishes the pronunciations of good and bad pronunciation with great accuracy. However, the system is not effective in differentiating those who have imperfect pronunciations. We strongly think that this can be improved with phoneme analysis and training that exhibits such characteristic.

References

1. McCrocklin S (2016) Pronunciation learner autonomy: the potential of automatic speech recognition. System 57:25–42
2. Foote J (2010) Second language Learners’ perceptions of their own recorded speech, Edmonton: PMC working paper series
3. Kruk M (2012) Using online resources in the development of learner autonomy and english pronunciation: the case of individual Learners. J Second Lang Teach Res 1(2):113–142
4. Barbosa F, Silva W (2015) Support vector machines, Mel-Frequency Cepstral coefficients and the discrete cosine transform applied on voice based biometric authentication. In: 2015 SAI intelligent systems conference (IntelliSys), pp 1032–1039
5. Neri A, Cucchiarini C, Strik W (2003) Automatic speech recognition for second language learning: how and why it actually works. In: International congress of phonetic sciences, pp 1157–1160. International congress of phonetic sciences, Barcelona
6. Hincks R (2003) Speech technologies for pronunciation feedback and evaluation. ReCALL 15
7. Gu L, Harris J (2003) SLAP: a system for the detection and correction of pronunciation for second language acquisition. In: International symposium on circuits and systems. Bangkok, pp 580–583

8. Practical Cryptography, http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/#eqn1. Accessed 31 Oct 2017
9. Du Y (2013) Biometrics. Pan Stanford Publishing Pte Ltd, Singapore
10. Recurrent neural networks tutorial part 1—introduction to RNNs (2017). http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/. Accessed 31 Oct 2017
11. Hansen J, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag 32:74–99
12. Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st international conference on machine learning, PMLR, pp 1764–1772
13. Chen S, Luo Y (2009) Speaker verification using MFCC and support vector machine. In: Proceedings of the International multiconference of engineers and computer scientists, pp 532–535. Proceedings of the international multiconference of engineers and computer scientists, Hong Kong (2009)
14. Downey A (2016) Think DSP. O'Reily Media
15. Probst K, Ke Y, Eskenazi M (2002) Enhancing foreign language tutors—In search of the golden speaker. Speech Commun 37:161–173
16. Rabiner L, Schafer R (2011) Theory and applications of digital speech processing. Pearson/Prentice Hall, Upper Saddle River [etc.]
17. Zhang F, Yin P (2009) A study of pronunciation problems of english learners in China. Asian Soc Sci 5
18. Moustroufas N, Digalakis V (2007) Automatic pronunciation evaluation of foreign speakers using unknown text. Comput Speech Lang 21:219–230

OpinionSeer: Text Visualization on Hotel Customer Reviews of Services and Physical Environment

Angela Siew Hoong Lee, Ka Leong Daniel Chong and Nicholas Chan Khin Whai

Abstract The popularity of online platforms such as social media, review websites and blogs in recent years has increased the volume of user-generated content. Travel websites such as TripAdvisor, Agoda, Booking.com has been a popular companion to many travellers when deciding which hotels to stay and where to go in a foreign country. Travel websites such as TripAdvisor, Agoda and Booking.com are sources for consumers to obtain past experiences, reviews and recommendations from other travellers who have visited a specific hotel. Thus, organizations and businesses now turn to big data analytics to inspire them to discover hidden insights and gain better understanding. In this paper, we study several factors which are impacting customer satisfaction and to identify various hidden insights towards travellers' experience through text mining.

Keywords Big data analytics · Text analytics · Hotel reviews
Customer satisfactions · Customer experience

1 Introduction

In the 21st century, every piece of information can be obtained through the internet. With the massive growth in the usage of the internet, user-generated contents have been on the rise as well especially in the online hotel booking websites. Travel websites such as TripAdvisor, Agoda and Booking.com are very important platform for travellers to share their experience and to find which hotel to stay in, restaurant to dine in and attraction places to visit when they are at the foreign place. The contents on such websites has been generated by travellers and their reviews has become so

A. S. H. Lee (✉) · N. C. Khin Whai
School of Science and Technology, Sunway University, Subang Jaya, Malaysia
e-mail: angelal@sunway.edu.my

K. L. Daniel Chong
School of Hospitality, Sunway University, Subang Jaya, Malaysia
e-mail: danielc@sunway.edu.my

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_35

important to other travellers. These websites are reliable and trustable sources for potential travellers to gather the necessary information such as past experiences, customer satisfaction, service quality of hotel, etc. Through these reviews, it influences the potential travellers choice of hotel patterns because if a hotel receives massive praises from past customers, it will make the first good impression towards potential customers who intend to visit the selected hotel, while if it is otherwise, it will make a poor first impression. As oppose to the traditional word-of-mouth recommendation, it takes ages for one to pass the information to another; however, with the convenience of the internet every customer has the capability to share their respective experiences to the world. Posting up their past experiences in the form of reviews will enable future customers to better understand the hotels' good and bad before making a reservation. In the era of the internet connected community, it has given every individual a platform to express and describe their experiences while allowing hotel managements to tap into these online sources to improve their brand, quality of service and reputation. This can result in providing travellers a better overall experience while ensuring they are satisfied.

The volume of user-generated content is mesmerizing and with these information, not knowing what lies behind it can potentially cause harm [1]. Because without tapping into these information, a brand would not know how customers have reviewed their products or services nor would they understand how customers feel towards their brand. When you do not know how your customers feel, you can be prepared to close. Similarly, in a hotel brand, detailed analysis of travellers reviews, enables the hotel management to understand your hotel from a third-person perspective, while ensuring you do not miss out on the "smallest-of-things" [1]. These will allow you to have an edge over your rivals and ensure you improve customer experiences, customer retention and customer loyalty [1]. Understanding customer reviews, customer expectations, demands and wants will allow you to identify the factors that affects customers satisfaction and enable hotels to create business processes to cater to the needs and wants of customers—thus, increasing potential edge over your competitors [1, 2].

Tourism Malaysia's mission statement would be to make Malaysia the prime contributor in the socio-economic development by driving excellence in the tourism industry [3]. Tourism Malaysia's objectives would be to create Malaysia to be an excellent tourist destination while they showcase the uniqueness, attractions and multi-cultural heritage of this nation and to increase the volume of tourists as well as extending the stay of tourists to drive overall tourism revenue [3].

1.1 Research Objectives

The objective of this study focuses on two key factors which may potentially have a major impact on customer satisfaction. These two factors are Services and Physical Environment. The hotel focused in this study is a famous hotel located in Klang Valley, Selangor. Data was collected solely from TripAdvisor as it is one of the leading online hotel review websites in the world. A recent article stated that TripAdvisor consists of more than 460 million reviews, making it the most popular online hotel review website [4]. In this paper, it focused on discovering hidden patterns, correlations and analysis on customer reviews to provide a better understanding of what do travellers feel or say about the chosen hotel. This will enable us to identify travellers demands and wants, travellers suggestions and maybe provide some suggestions to the operational management team of the hotel. By analysing these 2 key aspects it able to ensure that customer retention and customer loyalty of the hotel because by providing customers with excellent customer service such as ensuring staff are polite, helpful and friendly always or housekeeping staff excelling at cleaning and making up rooms will give customers a great impression towards the hotel and feel satisfied with the quality of service. In addition, with an excellent physical location it will grant travellers an ease of mind because every attraction is located within walking distances and it will ensure ease of access to various attractions. After some research, we identified some papers in the hotel industry regarding customer reviews, customer satisfactions and how customers feel towards your hotel [5, 6]. The freedom of speech has enabled every individual to write about their experiences and leave reviews regarding hotels they have visited and share it to the world [7]. SAS Text Miner is the primary tool which is used to perform pre-processing of data, and crawling of data to creating context diagram.

2 Literature Review

2.1 Measuring Hotel Guest Satisfactions

To measure hotel guests' satisfactions, researchers have begun to explore the dimensions of hotel guests' satisfaction. Disconfirmation paradigm and expectancy-value theories were used to explains hotel customer satisfaction noting that the influence of attribute on hotel guests' decision making is also significant [8]. Empirically, researchers have examined the relationship between the hotel features and hotel guests' satisfaction by considering the effects of customer perception [9, 10] and attitudinal loyalty [11, 12], as well as tourist, hotel, and service characteristics [13–15]. Despite substantial attention, the understanding of specific service quality and hotel features influence customer loyalty [16–18]. The effectiveness of service quality and hotel features in determining the satisfaction level among hotel guests remain uncertain [19, 20] and the past findings on the underwriting factors in increasing hotel guests' satisfaction need to be reconsidered [21] despite the latest studies have concluded that customer service and physical environment are the two most dominating aspects in deciding the degree of guests' satisfaction [22, 23].

2.2 Service Quality as a Satisfaction Driver

Service quality was widely hypothesized for being able to increase customer satisfaction and eventually improve customer retention [24–27]. The concept of service quality has then been a focus of great importance to service marketing researchers. It is mostly established that an affirmative association exists concerning service quality and customer satisfaction [24, 28]. The significant connection was realized by many researchers who observed the influence of perceived service quality on customer satisfaction in hospitality and financial services [24, 29]. Likewise, some studies have considered hospitality service consumers in specific and determined that service quality has a direct effect on consumer's satisfaction [30]. The same deduction was grasped by a study on the hospitality industry in China where the study discovered an important relation among service quality and satisfaction [31]. In spite of the uniformity of these discoveries, it should be considered that the focus of these studies was mainly on the examination of the cognitive constituent of satisfaction, with satisfaction being operationalized as an evaluative judgment. Rather limited studies have measured the "affective-emotional" element of hotel guest satisfaction [8, 28]. Among these studies that have inspected the association between perceived service quality and emotional satisfaction, [28] found that service quality in hotel services was positively related with emotional satisfaction. Other than that, similar study on service encounters in a hotel discovered that displays of emotion among hotel employees influenced hotel guests' evaluation of the joyfulness of hotel employees and that this perceived joyfulness mediated the relationship between the service employees' behavior and customers' own degree of satisfaction [32, 33].

2.3 Physical Environment as a Satisfaction Driver

The physical environment does play an important role in shaping the consumers' satisfaction. It was suggested that the strategic physical location and its environment of a hotel had a major influence on hotel guest's satisfaction and eventually improving intention to revisit the hotel [34]. Other than that, it was also reasoned that the physical environment of hotels could be efficiently applied to support the hotel brand positioning, to reposition the guest's perceptual mapping among competition, and to increase their guest satisfaction with the service encounter [35]. Past studies also confirmed the role of the physical environment of hospitality firms on enhancing satisfaction and stimulating purchase behaviors [36]. It was discovered that service provided by guest-contact personnel and physical environment had a considerably positive effect on hotel guest satisfaction and revisit intention [37]. A recent study argued that a hotel's physical environment has a direct link with cognitive responses, such as hotel guest beliefs and observations [38]. In the hotel context, the physical environment, such as strategic physical location, nearby environment and interior ambient condition provides first-time walk-in travelers with indications that convey the anticipated service quality and customer perceived value which will eventually increase their intentional purchase decision. Some studies have affirmed that a hotel's environment is a significant driver of a business traveler's perceived value and satisfaction [39, 40]. These studies confirmed the positive relationship between hotel location, customer perceived value and satisfaction. They examined the impact of hotel location in specific to convenience and nearby environment onto hotel guest's perceived value, satisfaction and behavioral intentions in the context of urban hotels. Outcomes indicated that location and its environment had substantial effects on travelers' perceived value. Furthermore, travelers perceived value also influenced travelers' satisfaction and future behavioral intentions. Not only was traveler satisfaction the utmost driver to future revisit intentions, but it also mediated the relationship between emotional responses and behavioral intentions when they stayed in the hotel [41].

2.4 Analytics and Hotel Industry

Big data analytics is the next frontier in changing the way hospitality industry should operates–be it improving customer experience or ensuring customer satisfaction or even enhancing business operations, analytics is the way to the future [42]. The potential behind big data analytics lies in the hands of the organization. Data alone is meaningless and useless, to make data valuable, it must be put to effective use. Therefore, it must be analysed to discover hidden patterns and trends, which could potentially improve quality of service, more effective and efficient business operations or even financial performance improvement. Data analytics can enable quicker and more accurate decision making. Looking at the hospitality industry which is providing service to millions of customers, where every individual customer having their own expectation and satisfaction level, the probability of hotels meeting every expectation and satisfaction is very slim [43]. Hence, hotels are turning to data analytics to run the show. Advanced analytical solutions give hotels the ability to understand and satisfy customers like never before [44]. An example of analytics assisting the hotel industry would be via marketing management—how to leverage on information to identify peak seasons, weather forecasts, local events, number of customers during a period. All this information can be analysed by analytics to help hotels alter their room prices while suggesting personalized promotional special campaigns to different customers [43]. An example of a hotel chain would be Marriot International Inc which is also one of the world's largest hospitality brand with 3,900 properties over 18 brands located at over 72 different countries [45]. Back in 2013, Marriot International Inc had a reported revenue figure of almost $13 billion by incorporating big data analytics into their business operations such as Group Pricing Optimizer (GPO) [45]. GPO is a price modulator for every individual statistically derived market segment to suggest and alter rooms rates based on demands; hence, it can replace the static rates with rates derived from advanced analytical techniques that has resulted in revenue gains for Marriot [45]. Potential customers can experience enhanced booking process speeds based on accurately tuned rates determined by dates and hotels. In addition, GPO provides specific answers to questions sales managers may have which enables almost instantaneous decision making [45].

2.5 Text Analytics

Listening to the voices of your customers is important and vital [46]. By capturing information on online review websites, it allows brands to better understand customer reviews regarding products or services. Today, the volume of reviews generated over the internet through social media or review websites are leaving organizations in chaos because of the inability to manage the ginormous volume of reviews [46]. Analysis of textual data gives us the ability to understand the context of a text and make decision more accurately. Text analytics can be defined as the translation of textual data into meaningful information [47, 2]. Text analytics focuses on extracting key pieces of textual contents from reviews, conversations and emails to discover meaningful information [47]. Through the understanding of textual contents, text analytics discovers the “who”, “where”, “when” of every conversation while providing information on the “what” and “buzz” of the topic of each conversation, “how” every individual involved in the conversation are feeling and “why” the conversation happens [47]. The potential to comprehend every part from the beginning to the end has been unlocked [47]. Thus, text analytics is key to unveiling hidden insights and knowledge to better understand customers and improve overall business [2].

3 Research Methodology

The popular websites for hotel booking or travelling sites are Airbnb, TripAdvisor, Agoda and Booking.com. TripAdvisor was used to carry out this research. We chose to do a hotel reviews on one of the famous hotel in Klang Valley, Selangor. We decided to perform a web crawl to find the factors that affects customer satisfaction in hotels. This is to allow a better understanding of the range of factors which affects customer satisfaction. Every individual and traveller have their own expectation and standard towards a hotel, therefore through their experiences you would be able to identify what you are doing right, what you are doing wrong, and how you could improve. TripAdvisor was selected for data collection due to the popularity it has within the travelling and hotel booking websites. The comments and reviews were collected from year 2015 until 2017, tallying up to a total of 2 years and the total reviews collected within this duration is approximately 3000 reviews. We uses Text Parsing, is to break down sentences into terms so any irrelevant text can be ignored. Text Filter, is a process whereby it is to identify terms and filter them according to the dictionary, repeated terms, stop-words, etc. Text Cluster, is the process of grouping similar terms and text together. Text Topic, is to group and combine terms into topic groups. Topic groups can only be formed if there are sufficient terms which are related or sufficient terms which are consistently talked/mentioned of.

4 Text Analytics

In this analysis we will be analysing based on 2 key areas which includes: Location and Service. Context diagrams will be generated for better understanding and an overview of key and important terms which are correlated. It allows us to identify key patterns and trends from the important linkages and terms. Concept Linking Diagram shows the links a word has with other terms. From this diagram, we can identify the various terms mentioned in relation to the term we are analysing. Through the analysis of Concept Diagram, it allows data analysts to tell a better story and provide justification as to why certain terms are mentioned and what association/relationship it may have with other terms. The first term we analysed was Service (Fig. 1).

We observe that customers had a memorable experience because of the staff and concierge services, which could indicate that many customers are very happy with the reception staff. This is very important because the first impression given to customers leaves the strongest memory and impression of the hotel. This conclusion can be pre-justified by looking at the concept linkage in "Memorable", when we expanded the link into "Concierge", we can see that travellers post comments such as simple and friendly which states that the front desk personnel left a good strong impression. It also shows that the travellers had a great stay, it was enjoyable, pleasant, enjoyable stay, etc. This shows that many travellers to that particular hotel were left with a smile on their faces with the wonderful experiences they had.

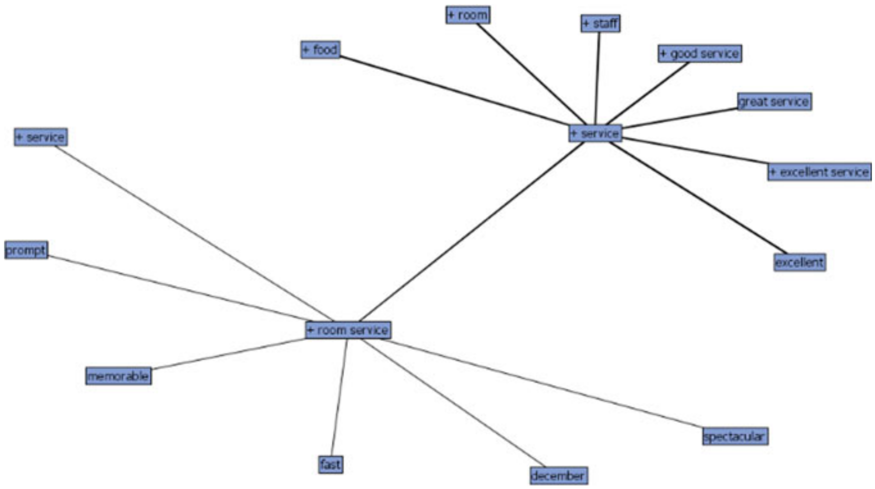


Fig. 1 Context diagram (Service—Room Service)

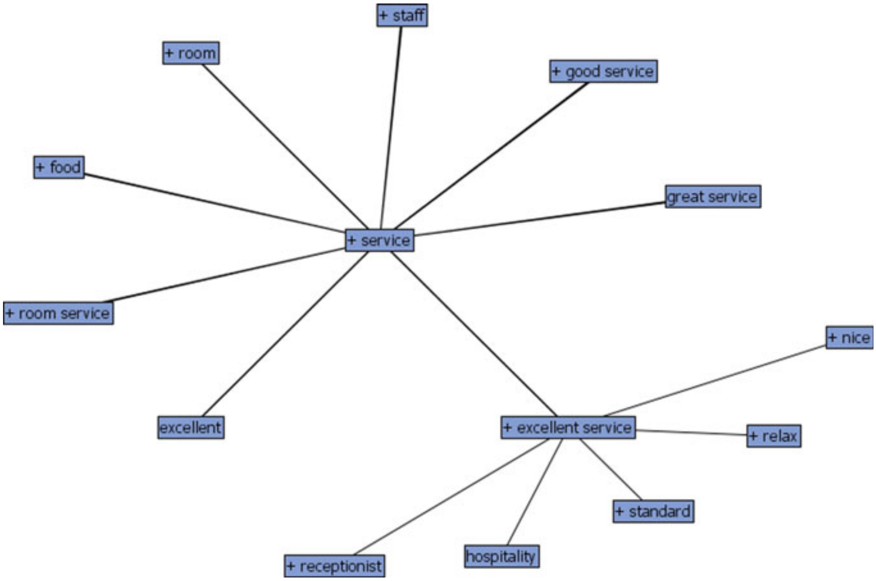


Fig. 2 Context diagram (Service—Excellent Service)

Figure 2 is another link which caught our attention which is Excellent Service. Upon expanding the links, we can identify various terms connected such as: “Receptionist, Hospitality, Standard, Relax and Nice”. This simply justify more on the excellent services provided by the hotel to the travellers.

Figure 3 shows the context diagram of Location. In this section, we want to understand and identify if location is a key aspect when hotel customers choose a hotel and understand how much of an impact does location have. Above is a context diagram to show the terms link to Location. Looking into more detail, we decided to expand the term Mall—it showed 5 links which are connected when mall is mentioned. The 5 terms are “Park, Shop, Restaurant, Huge and Connect”. From these terms, we can roughly have an idea of what customers’ mention when they talked about location. The most prominent aspect would be the shopping mall next to the hotel. Mall has strong connection with “Shop, Restaurant, Huge and Connect”. This can explain that the mall is connected to the hotel and that the mall has many shops, ice skating ring, cinema as well as the mall being very huge and it has many restaurants to choose from. These are some of the key patterns which has been identified. This clearly shows the importance of a strategically located hotel that has provides convenience and accessibility to the customers.

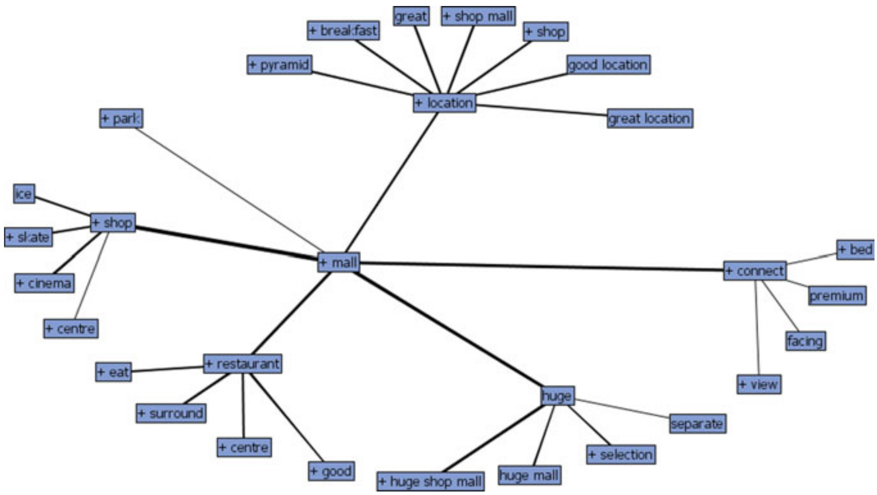


Fig. 3 Context diagram (Location—Mall)

5 Discussion

From these analysis, we managed to gather meaningful insights and determine underlying patterns and trends. First and foremost, looking at the Location ratings, we can conclude that the chosen hotel in Klang Valley is a cut above its rivals. Location of the hotel place an important role and this has shown from the reviews we have gathered from the travel websites. On the other hand, Service clearly draws the conclusion of whether travellers would have a memorable and unforgettable experience. Under Service, we manage to identify that Room Service, Memorable, Good Service, Great Service, etc. clearly shows how leaving a good impression from the start will make the day of the customers. Customers have stated how prompt, fast and responsive Room Services have clearly enhanced their experiences. And having good Service, have given customers a memorable experience and improve their overall stay. Customers left comments such as enjoyable stay, pleasant stay, etc., which shows how Service has a major impact on customers stay and their experiences. Ensuring customers are served well and treated well is a vital and crucial part for a hospitality industry, therefore when customers leave comments such as excellent service, great service, etc. it definitely shows that the hotel strongly instil the principle that customers are always the priority.

6 Conclusion

Conceivably the most significant implication of this study is the importance of service quality and location and its environment in deciding the degree of hotel guests' overall satisfaction. Hotel managers need to properly benchmark the degree of their service offerings and its detail of quality, given that hotel guests' perceptions and satisfaction of a hotel property not merely have direct impact on their perception towards the hotel performance, but also moderate the effectiveness of the hotel in driving greater revisit intention. In view of this perspective, the hotel should continue to improve and maintain their standard of service. Eventhough there have been rare reviews posted by customers stating that the hotel offers excellent and great service, they must not stop there but continually provide customers with unforgettable experiences. Because customer service is vital in all hospitality industry which heavily rely on personal interaction [48]. Other than that, this study indicates that a more competitive location and its environment stand to gain more consumers' satisfaction than less strategic locations. For less strategic hotel locations, the perceived value of the hotel service offerings may not have a complete positive influences on overall hotel guets' satisfaction.

In view of the importance of service quality and location in potentially affecting consumers' evaluation and satisfaction, hotel managers should uphold an importance on improving service perceptions and making sure that all ranges of hotel services (Rooms, Housekeeping, Food and Beverage, Check-in etc.) offers solid value which focuses on guests' special needs and wants. The results of this study propose that the provision of a highly valued hotel services can improve hotel guests perceived value and experience. This is true to hotel location and what surround the hotel. For strategically located hotels, its not only impact on the hotel guest purchase behavior under certain conditions, but can also strengthen consumer attitudes toward the overall hotel experience. With this in mind, our results suggest two tactics. Hotels need to focus on distinguishing their service offerings from competiting hotels and also on increasing the accessibility and visibility fo their premises. One way to upgrade service value and evaluation is to position a hotel customer service program as being a privilege [49]. Hotel managers must also put efforts in providing meaningful and customized services that match their context and capacity. These modest alterations can lead to growths in both hotel consumer satisfaction and referral behavior.

In terms of improving the competitiveness of hotel existing location, there is a need for better differentiation of location accessibility and attractiveness to ensure that first-time visitors to recognize the convenience, functional and emotional value that is offered by a particular location relative to the competition. Just as a new hotel would never position their premises without considering its distance from transportation hub, business district, tourist sites, eateries and competitors. Location elements that might realize this include hotel, airport, business centre and city centre connectivity for in-house guests; proper site inspection for new hotel developer considering the potential retails which add value to hotel guests.

In conlusion, this study not only explored data from one hotel, there is a need to further study to determine whether the discoveries of this research apply to other form of hotel settings considering their difference in business nature, service features, and landscape. Additionally, a thorough qualitative analysis could be done to develop a more direct measure between the types of reviews such as moderate reviews and extreme reviews.

References

1. McGarrity L (2016) What sentiment analysis can do for your brand. Marketing Profs, 5 Apr 2016. http://www.marketingprofs.com/opinions/2016/29673/what-sentiment-analysis-can-do-for-your-brand. Accessed 11 Apr 2017
2. Hoong ALS, Lim TM, Leow SK, Aun JLR (2012) A study on the use of "Yams" for enterprise knowledge sharing. In: Digital Information and Communication Technology and it's Applications (DICTAP), 2012 Second International Conference on (pp. 183–188). IEEE
3. Malaysia T (2017) About tourism malaysia, tourism malaysia. http://www.tourism.gov.my/about-us/about-tourism-malaysia. Accessed 11 Aug 2017
4. Salzman A (2017) Why TripAdvisor is getting a bad review, Barron's next, 16 Feb 2017. http://www.barrons.com/articles/why-tripadvisor-is-getting-a-bad-review-1487266932. Accessed 14 Apr 2017
5. Darini M, Khozaei F (2016) The study of factors affecting customer's satisfaction with three star hotels in Dubai. Int J Adv Eng Manag Sci 2(2):21–24
6. Xiang Z, Schwartz Z, Jr JHG, Uysal M (2015) What can big data and text analytics tell us about hotel guestexperience and satisfaction? Int J Hosp Manag 44:120–130
7. Byng R (2015) 15 Years of TripAdvisor: how 200 million reviews changed the way we travel, independent UK, 25 Feb 2015. http://www.independent.co.uk/travel/15-years-of-tripadvisor-how-200-million-reviews-changed-the-way-we-travel-10070175.html. Accessed 14 Apr 2017
8. Barsky JD (1992) Customer satisfaction in the hotel industry: meaning and measurement. Hosp Res J 16(1):51–73
9. Kandampully J, Dwi S (2000) Customer loyalty in the hotel industry: the role of customer satisfaction and image. Int J Contemp Hosp Manag 12(6):346–351
10. Castro CB, Enrique MA, David MR (2007) The influence of market heterogeneity on the relationship between a destination's image and tourists' future behaviour. Tour Manag 28(1):175–187
11. Bowen JT, Shiang-Lih C (2001) The relationship between customer loyalty and customer satisfaction. Int J Contemp Hosp Manag 13(5):213–217
12. Nam J, Yuksel E, Whyatt G (2011) Brand equity, brand loyalty and consumer satisfaction. Ann Tour Res 38(3):1009–1030
13. Shankar V, Smith AK, Arvind R (2003) Customer satisfaction and loyalty in online and offline environments. Int J Res Mark 20(2):153–175 (2003)
14. Hu HH, Jay K, Thanika DJ (2009) Relationships and impacts of service quality, perceived value, customer satisfaction, and image: an empirical study. Serv Ind J 29(2):111–125
15. Chand M (2010) The impact of HRM practices on service quality, customer satisfaction and performance in the Indian hotel industry. Int J Hum Resour Manag 21(4):551–566
16. Back KJ, Parks SC (2003) A brand loyalty model involving cognitive, affective, and conative brand loyalty and customer satisfaction. J Hosp Tour Res 27(4):419–435
17. Lee J, Graefe AR, Burns RC (2007) Examining the antecedents of destination loyalty in a forest setting. Leis Sci 29(5):463–481
18. Gross MJ, Graham B (2008) An empirical structural model of tourists and places: Progressing involvement and place attachment into tourism. Tour Manag 29(6):1141–1151

19. Rahimi R, Metin K (2017) Impact of customer relationship management on customer satisfaction: The case of a budget hotel chain. J Travel Tour Mark 34(1):40–51
20. Pozo H, Sergio Luiz DAM, Takeshy T (2016) Hospitality practices as sustainable development: an empirical study of their impact on hotel customer satisfaction. Tour Manag Stud 12(1):212–222 (2016)
21. Pizam AAP, Valeriya S, Valeriya S, Taylor E, Taylor E (2016) Customer satisfaction and its measurement in hospitality enterprises: a revisit and update. Int J Contemp Hosp Manag 28(1):2–35
22. Guo Y, Barnes SJ, Qiong J (2017) Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. Tour Manag 59(1):467–483
23. Subramanian N, Angappa G, Yanan G (2016) Innovative service satisfaction and customer promotion behaviour in the Chinese budget hotel: an empirical study. Int J Prod Econ 171(1):201–210
24. Cronin JJ, Brady MK, Tomas MHG (2000) Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. J Retail 76(2):193–218
25. Duncan E, Greg E (2002) Customer service quality and financial performance among Australian retail financial institutions. J Financ Serv Mark 7(1):24–41
26. Janda S, Trocchia PJ, Kevin PG (2002) Consumer perceptions of Internet retail service quality. Int J Serv Ind Manag 13(5):412–431
27. Kang GD, Jeffrey J (2004) Service quality dimensions: an examination of Grönroos's service quality model. Manag Serv Qual Int J 14(4):266–277
28. Ladhari R (2009) Service quality, emotional satisfaction, and behavioural intentions: A study in the hotel industry. Manag Serv Qual Int J 19(3):308–331
29. Cronin Jr JJ, Taylor SA (1992) Measuring service quality: a reexamination and extension. J Mark 56(3):55–68
30. MK Brady, Knight GA, Cronin Jr JJ, Tomas G, Hult M, Keillor BD (2005) Removing the contextual lens: a multinational, multi-setting comparison of service evaluation models. J Retail 81(3):215–230
31. He Y, Wenli L, Kin KL (2011) Service climate, employee commitment and customer satisfaction: evidence from the hospitality industry in China. Int J Contemp Hosp Manag 23(5):592–607
32. Barsky J, Leonard N (2002) Evoking emotion: affective keys to hotel loyalty. Cornell Hotel Restaur Adm Quart 43(1):39–46
33. Söderlund M, Sara R (2004) Dismantling positive affect and its effects on customer satisfaction: an empirical examination of customer joy in a service encounter. J Consum Satisf Dissatisfaction Complain Behav 17(1):27–41
34. Booms BH, Bitner MJ (1982) Marketing services by managing the environment. Cornell Hotel Restaur Adm Quart 23(1):35–40
35. Ali F, Muslim A, Kisang R (2016) The role of physical environment, price perceptions, and consumption emotions in developing customer satisfaction in Chinese resort hotels. J Qual Assur Hosp Tour 17(1):45–70
36. Ryu K, Hye-Rin L, Woo Gon K (2012) The influence of the quality of the physical environment, food, and service on restaurant image, customer perceived value, customer satisfaction, and behavioral intentions. Int J Contemp Hosp Manag 24(2):200–223
37. Han H, Sunghyup SH (2017) Impact of hotel-restaurant image and quality of physical-environment, service, and food on satisfaction and intention. Int J Hosp Manag 63(1):82–92
38. Chang FS, Ishak N, Ramly ASM, Ramlan NS, Chu CH (2016) The effect of physical environment on behavioral intention through customer satisfaction: A case of five-star beach resorts in Langkawi Island, Malaysia. In: Heritage, culture and society: research agenda and best practices in the hospitality and tourism industry, Bandung, Indonesia
39. Mattila AS, Cathy AE (2002) The role of emotions in service encounters. J Serv Res 4(4):268–277
40. Yang Y, Zhenxing M, Jingyin T (2017) Understanding guest satisfaction with urban hotel location. J Travel Res 1:1–17

41. Kandampully J, Dwi S (2003) The role of customer satisfaction and image in gaining customer loyalty in the hotel industry. J Hosp Leis Mark 10(1):3–25
42. Shipley M (2017) Hospitality big data analytics, Fourth, 24 Mar 2017. https://www.fourth.com/en-gb/blog/hospitality-big-data-analytics. Accessed 23 May 2017
43. Mishra S (2016) Big data analytics in hospitality industry, 7 June 2016. https://www.linkedin.com/pulse/big-data-analytics-hospitality-industry-sandeep-mishra. Accessed 15 Apr 2017
44. Marr B (2016) How big data and analytics are changing hotels and the hospitality industry, Forbes, 26 Jan 2016. https://www.forbes.com/sites/bernardmarr/2016/01/26/how-big-data-and-analytics-changing-hotels-and-the-hospitality-industry/#3d66605f1c22. Accessed 15 Apr 2017
45. Gupta B (2017) How analytics can help the hospitality industry?, Analytics India Magazine, 11 Apr 2015. http://analyticsindiamag.com/how-analytics-can-help-the-hospitality-industry/. Accessed 23 May 2017
46. Kho ND (2010) Customer experience and sentiment analysis. KM World, pp 10–20
47. Gutierrez D (2015) Text analytics: the next generation of big data, inside big data, 5 June 2015. http://insidebigdata.com/2015/06/05/text-analytics-the-next-generation-of-big-data/. Accessed 20 Apr 2017
48. Yongchaitrakool S (2014) The effect of customer expectation, customer experience and customer price perception on customer satisfaction in hotel industry. In: International conference on management science, innovation, and technology, pp 66–74
49. Winer RS (2001) A framework for customer relationship management. California Manag Rev 43(4):89–105

Comparisons in Drinking Water Systems Using K-Means and A-Priori to Find Pathogenic Bacteria Genera

Tevin Moodley and Dustin van der Haar

Abstract As water resources have become limited, there have been increased cases in illnesses related to waterborne pathogens, with this is mind studies and investigation needs to be done on alternative water sources such as, ground water and common water sources such as surface waters, to ensure that water provided to consumers are safe to consume. This research paper compares bacterial genera in both ground and surface source waters for drinking water systems, based on 16S rRNA profiling using machine learning methods, such as K-means and A priori. 16S can be used to identify and differentiate between bacterial genera. Not only is it important to identify specific bacterial genera found in water sources, but the relative abundance needs to be examined to determine whether groundwater is a more viable drinking water source than surface water. Using recent incidences of water-borne illnesses that have been reported across South Africa, five key bacterial indicators to determine water quality and safety can be identified, which can be found in both groundwater and surface waters. Captured data from samples collected is used to determine the abundance of each bacterium for each water sample in a more efficient and effective manner the five indicators outlined for this project are; *E. coli* (*Escherichia*), *Legionella*, *Hemophilia*, *Bdellovibrio*, *Streptococcus*. The dataset, used contained bacterium from both ground and surface waters using dimensional techniques and many parameters can be reduced for more efficient processing. The algorithms used include K-Means to cluster the data to allow for interpretation, A Priori algorithm to get the frequent items so that association rules can be derived, which allows patterns to be realized and SVM (support vector machine) to predict the error of new data coming into a stream. Using the results produced by the algorithms, it was discovered

T. Moodley · D. van der Haar (✉)
Academy of Computer Science and Software Engineering, University of Johannesburg, Cnr University Road and Kingsway Avenue, APK Campus, Johannesburg 2006, South Africa
e-mail: dvanderhaar@uj.ac.za

T. Moodley
e-mail: tevinmoodley7@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_36

that the mean relative abundance of the pathogenic organisms found in groundwater was higher than that found in surface water. Results indicated that automated, scalable water viability assessment is feasible using the methods proposed, which make it an attractive avenue of research as the Internet of Things (IoT) in this domain develops.

Keywords Bacteria · Water assessment · Hadoop · PCA · K-means · A priori SVM

1 Introduction

Over time South Africa's concerns over water resources have increased due to the limited drinking water sources, the importance of considering groundwater and surface water as viable drinking water sources has increased, especially in rural areas where water filtration and treatment is not done efficiently. Groundwater is the water found underground between the cracks and spaces in the soil, sand and rocks. Once stored it moves slowly through geological formations of aquifers (sand, soil and rock). Whereas surface water is water found in rivers, lakes, dams and oceans. Water is the most abundant resource available on earth, water covers approximately 70 percent of the earth [1]. However, the issue is not relative to the amount of water available but rather on the viability of waters available. Bacterial composition influences water quality and can have a limiting effect of which water sources are used for drinking water systems. Water is essential for the survival of individuals as well as the functioning and economic activities of communities. Access to water is therefore widely considered to be a basic human right [2]. Chemical quality of drinking water is closely monitored however, the microbial quality of water needs to be safe for consumption as it is estimated that in South Africa, 1.8 million people die each year from water-borne diseases [2]. The challenge with providing good quality and safe drinking water has been made more complex with human contributions and environmental changes. Microbes found in drinking water at low abundance does not pose a health risk, thus microbial abundance needs to be monitored [3]. In this research paper there are examples of how water-borne bacteria cause illness among humans. The paper will discuss the five key bacterial pathogens that are most commonly known for water-borne illnesses. The data collected from the different water sources will be stored on a distributed file system so that it may be processed at a more efficient and effective rate. Using Apache Hadoop [4] and other big data tools, the data will be scaled down to minimize the processing time. A discussion of how the data is analyzed will be done to obtain results which can then be used to make a conclusion as to which water source is more of a viable option for drinking.

2 Problem Background

It has been noted by the Water Supply and Sanitation in South Africa, that water pollution has increased over the last decade. Some of the biggest contributions to water pollution include; urbanization, physical disturbance of land due to construction, inadequate sewage collection and treatment processes and the increase in fertilizers for agricultural reasons [5]. The city of Cape Town in South Africa has seen a dramatic decline of water availability over the recent years, it has been reported that the last 10% of dam water is unusable due to increased water-borne illnesses such as gastroenteritis and diarrhea due to high levels of pollution [6]. For that reason, the monitoring and removal of water-borne bacteria from surface water and groundwater had to be investigated, Mwabi et al. [7] conducted experiments to determine the most cost-effective way to remove water-borne diseases from the two different water sources, based on their findings it was concluded that the filtration was only cost-effective in rural communities [7]. Their findings suggested that bacteria are more likely to be prominent in rural areas where there are lakes and rivers due to the water sources being more exposed to external environmental factors, contamination is more likely. However, groundwaters as an alternative source needs to be further examined as the bacterial abundance and composition are unknown, since filtration is done on such a large scale. The need for disinfection and treatment requirements will be reviled in determining the quality of water sources and the cost implications associated.

The UN environment program ranked South African water 47th out of 122 countries, and in 2012 it showed that there was an 8% drop in water quality. These trends make it abundantly clear that water quality has dropped [8]. These incidences have raised many issues. Thus, identification of pathogenic bacteria in these water sources will be informative when choosing a source water for drinking water systems [1]. Contrary to expectations of groundwater and surface water being two separate entities, Winter and colleagues [9] wrote a report governed by the US government to prove that groundwater and surface water are in fact one resource, they proved that the development of either one of these resources affects the quantity and quality of the other due to interconnections between them. But upon their research, they were unable to determine the bacterial implications between the two sources which would imply that there are differences between the two in terms of associated water-borne diseases [9]. Knowing that there are differences between groundwater and surface water microbes, it can be concluded that a change in environment would lead to differences in bacterial community. Thus, the bacterium in the different water sources needs to be analyzed to identify which water source could potentially be more dangerous along with a way to process this data efficiently in a scaled manner so that assumptions can be made faster thus, allowing one to prevent illnesses among humans as well as identifying a potential alternative water source.

2.1 Bacteria to Be Analyzed

An average a glass of drinking water contains around 10 million bacteria [10], so it stands to reason that with this high bacterial load present, there are some bacteria that will cause illness among your average human being. A select number of bacteria are being investigated in the study that is known to cause illness among consumers five different bacteria will be examined from four different water samples. *E. coli* (*Escherichia*), this is a bacterium that lives inside the gastro-intestinal tracts of humans and animals [11, 12]. Symptoms include diarrhea. *Legionella*, this bacterium is best known for causing the Legionnaire's disease (Pneumonia), the bacterium is largely aquatic and can thrive in areas such as rivers, lakes and oceans [13]. Symptoms include fever, cough, chills, muscle aches. *Haemophilus*, there are several types which can cause different types of illness involving breathing, body structure and nervous system [14]. Symptoms, flu-like symptoms. *Bdellovibrio*, this bacterium is found in man-made habitats, generally on the roots of plants and in the soil which leaches into the water [12]. Symptoms include damages to the human gut. *Streptococcus,* this bacterium is known for infecting animals, including humans, with associated diseases ranging from strep throat to necrotizing fasciitis [10]. Symptoms include infections in parts of the body.

3 Experiment Setup

The aim of the project is to compare different water sources and identify bacterial genera and their abundances using machine learning algorithms more notably A Priori and K-Means clustering.

3.1 Water Sampling and Data Scaling Down Process

Sterile containers will be used for collection of water samples. Sampling was conducted, there were 10 groundwater samples for Ground1 and Ground2 (G1 and G2) and 12 surface water samples were taken, Surface1 (S1) and Surface2 (S2) respectively, the sampling [15] was collected by collaborating with "anonymous", the location of the water sources cannot be disclosed. 8–16 L of water was collected, the water samples collected from these 4 sources were stored in a −20° Celsius fridge until DNA extraction was performed [16]. Once the DNA sequence was interpreted into readable data, the dataset consisted of 165171 rows along with 46 columns. Each row represents a different OTU number which gives identity to each genus, bacteria can be from the same genus but could have a different species identification, which would then be represented by a different OTU number. All files were scaled down

Table 1 The mean relative abundance of 5 different bacteria across all four sample points

Genus	mra_G1	mra_G2	mra_S1	mra_S2	Total average
Bdellovibrio	0.000105	0.001695	0.000217	0.000176	0.0548
Escherichia	0.010496	0.001895	0.000008	0.000627	0.3257
Haemophilus	0.000036	0.000133	0	0.000004	0.0043
Legionella	0.000093	0.001294	0.00013	0.000358	0.0469
Streptococcus	0.024059	0.001487	0	0.000104	0.6413

and made using HIVESQL queries which use the HDFS [17] and R Studio [18] to process the data faster and efficiently.

4 Results

Using A Priori to discover association rules can be challenging if the support threshold is low, as the association then becomes futile. K-Means clustering had few limitations but the biggest challenge faced was establishing the axis of best fit but was later made easier using Orange3. Results will be illustrated to help better understand the data. Table 1, a sample set and the final results indicate the five key bacteria that was outlined in the problem background. Each genus was averaged out in their respective sample points. The table shows the mean relative abundance for samples G1, G2, S1 and S2.

Figure 1 is a representation of a PCA (principal component analysis) which forms clusters based on correlations of the sample points to one another. The clusters are grouped based on their similarity of each genus to one another, this could be informative of the bacterial that is dominant in the system. Any data that falls outside the main cluster between 0 and 0.5 is a threat, the bacterium that will appear outside the main cluster will be Flavobacterium and Unclassified.

Figure 2 is a 100% column stacked graph, Table 1 was taken and graphically represented as indicated. Ground 1 sample shows streptococcus and Escherichia have the highest relative abundance, ground sample 2 shows an even distribution of all the bacteria apart from Haemophilus, surface sample 1 is predominantly made up of Bdellovibrio and Legionella and surface sample 2 exhibits high amounts of Escherichia.

5 Discussion and Critique

The K-means algorithm was used along with Scikit-learn library to obtain Fig. 1. Scikit-learn is the primary machine learning library that is used in python to implement most algorithms [19], in this case (SVM) support vector machine. The K-means

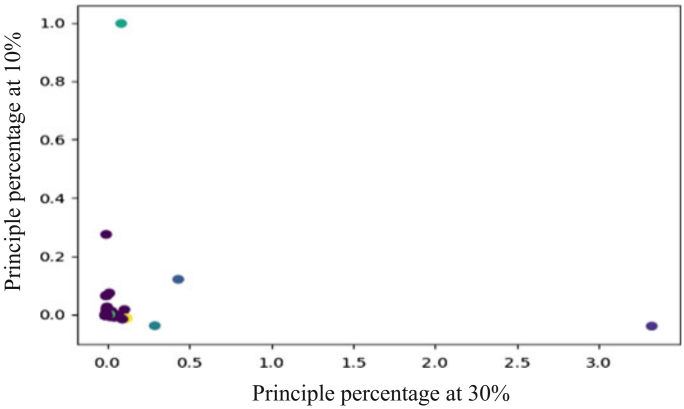


Fig. 1 PCA clustering based on the genus of bacteria and their abundance values

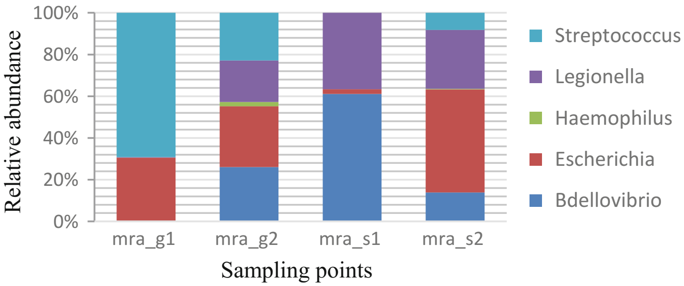


Fig. 2 Graphical representation of the mean relative abundance of 5 different bacteria across all four sample points

model was initialized to 8 clusters, it is set to random state 1, which is used to reproduce the results at a later stage. The algorithm was made to select only numeric values, which removed the genus column and the sample(OTU) column since clustering does not allow for it. Since data that is higher than three dimensions is outside the realm of human understanding and physics [19], dimensionality reduction was applied using the principal component analysis(PCA), it allowed multiple columns to be turned into fewer columns. From Fig. 1, clusters of bacteria can be seen and this discovers the axis of the data, the axis will enable the program to see the genus to abundance concept, therefore correlations between each cluster of bacteria can be made. Each cluster is color shaded to help distinguish the differences, thus allowing analysts to dive more into which bacteria are in each cluster to learn more about what factors, if there are, can cause bacteria to be clustered (i.e. associated factors such as environmental influence). Interpretation of the graph using the abundance of each bacteria across each month, a genus with the similar abundance will be grouped together and most of the bacteria range from 0 to 0.5%, this cluster can be ignored, microbiologists should be concerned with the other clusters that fall outside of the range as those are the bacteria that exhibit the highest abundances relative to the sample points.

A prediction of the total number of each genus can be done, to do this, the correlation between the total and the other columns using the final data set is used, using Table 1 column 6 is compared to columns 2–5, which will show what column will predict the total reading the best. The correlation method on Pandas is used to allow this, executing it through the HDFS the following results were obtained; mra_G1-0.938377, mra_G2-0.967416, mra_S1-0.933359, mra_S2-0.966411, total-1.000000. It is noticed that that mra_G2 correlates best with the total column. Mra_G2 indicates that the abundance for each genus in that sample set mostly determines the total abundance of each bacteria, so mra_g2 is the biggest contributor for the total reading, possibly the sample set taken from ground 2 had readings for every genus and it exhibited the highest readings, these are the types of conclusions that can be made. In our case, the values are extremely similar, the chances of the total correlating to one other column are small.

Since an interesting correlation was made between mra_G2 and the total, the data can be split into training sets and test sets, overfitting is evaluating an algorithm on the same data [19], to avoid this the algorithm will be trained on a set consisting of 70% of the data and the remaining 30% will be used for testing. The results that are obtained after execution are (493, 6), (123, 6), this ensures that the training set and test set data is kept separate. Linear regression is then used to predict error since, our target value and predictor values are linearly correlated, once the data is trained the prediction error can be determined, which will give the average error for each prediction made. The prediction error that was given is 0.0107004938911, which means for every new genus (bacteria) read into the data set the mra_G2 column values will be off by an average of 0.0107004938911, which indicates the prediction that the mra_G2 value will determine the total value is very interesting.

Using Table 1 and Fig. 2 the results can further be discussed. *Bdellovibrio* is significantly low in abundance, the highest recorded reading was taken from ground sample 2, the total average reading is insignificant which means that there is no risk to humans due to the low dosage that would be consumed. Interestingly *Bdellovibrio* appears to be more prominent in groundwater rather than surface water, however, this can be expected since *Bdellovibrio* thrives in closed environments, that contain plants and soil [13]. *E.coli (Escherichia)*, has a noteworthy reading in ground sample 1 of 0.010496%, the dosage consumption is harmless to humans, this result is surprising as it was unexpected since, *E.coli* is known to thrive in rivers, lakes and oceans, surface water sources, as it is caused by fecal matter which contaminates open water environments. An explanation of this finding could be due to infiltration of surface waters or other contaminants into the groundwater sources [20], this finding opens questions to further studies and understanding of the bacterial community in groundwaters, especially if these sources are used for drinking water systems. The total average of *E.coli* is 0.3257% across all sample points, this value is low and not harmful. *Haemophilus*, although seen in higher abundances in groundwater showed low insignificant values across all sample points and should not be considered a high-risk influential bacterium for water-borne diseases. *Legionella* readings in the groundwater samples were higher than that of the surface water samples, which is extremely unusual as *Legionella* thrives in surface water sources [13] due to their favorable bacterial environments, likewise exhibited with *E.coli*. The total average reading of 0.0469 which is harmless to human. *Streptococcus* is seen in higher abundances in groundwater rather than surface water.

6 Conclusion

Using online machine algorithms along with Big data mining techniques the relative abundances of the 5 key bacteria were analyzed to determine which is a more viable drinking water source between groundwater or surface water. The system accommodates for massive datasets and process it in a manner of a few hours using Apache Hadoop [17] will allow data analysts to obtain results faster thus making assumption quicker and more efficient. The system is dynamic in the sense that any number of water samples can be taken and compared to one another, this creates an automated system that requires less input from a data analyst, thus assisting them in obtaining more accurate results. From the results, the relative abundance of bacterial genera in groundwater sources showed to be higher to that of surface water sources, which was not expected since surface water sources are prone to water-borne diseases as they are more exposed to contamination. Winter [9] studies show that surface water and groundwater are in fact one source, this could be the reasoning for groundwater exhibiting higher abundance reads than surface waters. Additional factors to consider in the evaluation of the bacterial community could be chemical parameters as they can be influential on the bacteria present. However, although groundwater showed higher abundances all sample points for both groundwater and surface water showed low insignificant dosage amounts for any harm to the consumers. Mwabi et al. [7] concluded that filtration of water was needed for surface waters, however, perhaps groundwaters after the above finding should in fact also be filtered as a precaution. The research that was conducted was subjected to the water sources. Further water quality tests are important for all drinking water. External environmental factors should be addressed and understood when interpreting water systems such as weather implications and temperature, understanding this would also lead to better predictive indicators, saving money and disinfection and treatment processes to tackle future challenges.

References

1. Hordon RM (2005) Water and surface hydrology. Water Encycl 7:89–112
2. Nkanyani G (2014) Water pollution. Rand Water South Africa 6:98–200
3. Serra LJ (2017) Tap water is making us sick. Daily Voice 1:45–56
4. Mike C (2016) Hadoop ecosystem overview. Big Data Blog 2:9–10
5. Taylor T (2016) South Africa tap water np cleaned properly. Experts 1:98–145

6. Colby D (2015) Life cycle of malaria parasites. Med Web 9:45–46
7. Mwabi JK, Mamba BB, Momba MN (2013) Removal of waterborne bacteria from surface water and groundwater by cost effect household water treatment. South Africa Publ Serv 2:65–99
8. Liberatore S (2015) Fancy a drink. Sci Technol 5:2–5
9. Winter TC, Harvey JW, Frank LO, Alley WM (1998) Groundwater and surface water. Res Gate 2:45–55
10. National research council (2004) Waterborne pathogens, indicators for waterborne pathogens. National Academy press, (1), pp 550–800
11. Agee J (1975) Protecting drinking water, responsibilities under the safe drinking water act. Environ Protect Agency 9:45–105
12. Parte AC (2001) E coli. Sci Direct 1:45–62
13. Markelova N (2010) Predacious bacteria, Bdellovibrio with potential for biocontrol. Int J Hyg Environ Health 6:428–431
14. MeilinePlus (2001) Drinking water. Mediline Plus (5) 44–45
15. World Health Organization (2016) Water sampling. World Health Organ 40:1–6
16. Science Learning Lab (2009) DNA extraction. Sci Learn Lab 1:45–99
17. Wiley J (2015) Data science and big data analytics: discovering analyzing, visualizing and presenting data-Hadoop. Res Gate 1:300–325
18. Maindonald JH (2008) Using R for data analysis. Math Centre 6:98–110
19. Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive data sets. Stand Univ 1:560–750
20. Game T (2017) Fountains and springs. Ground Water Found 2:89–99

Optimal Lightweight Cryptography Algorithm for Environmental Monitoring Service Based on IoT

Jongmun Jeong, Larsson Bajracharya and Mintae Hwang

Abstract Even though IoT devices for monitoring and managing the environment are already present at national level, vulnerability related to security in these systems still prevails at large. Therefore, it is required to study and implement a feasible security protocol in these systems. In this paper, we tested optimal lightweight cryptography for security enhancement in an IoT based Environmental Monitoring System that monitors and controls the surrounding temperature and humidity. We compared the CPU usage and processing time when various lightweight cryptography was applied. This helped us to conclude that the optimal lightweight cryptography for IoT based system is LEA which is a block cipher.

Keywords Lightweight cryptography · Security · IoT · Monitoring
Remote control

1 Introduction

Ubiquitous, which uses a variety of information technologies regardless of time and place, is easily realized due to its high-performance mobile computing device and is commonly used today as a smart phone. Since 2009AD, the number of smart phone users have gone high drastically. Similarly, IoT (Internet of Things) has also been recognized as having a high possibility of success. In fact, the IoT based devices that

J. Jeong (✉) · L. Bajracharya
Department of Eco-Friendly Offshore Plant FEED Engineering, Graduate School, Changwon National University, Changwon 641-773, Korea
e-mail: jhs7986@gmail.com

L. Bajracharya
e-mail: larssonbajra@gmail.com

M. Hwang
Department of Information & Communications Engineering, Changwon National University, Changwon 641-773, Korea
e-mail: professorhwang@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_37

Table 1 The number of malware targeting IoT samples detected each year

Year	2013	2014	2015	2016	2017
Malware targeting IoT	46	193	696	3219	7242

were being used in 1992 AD were around 1 million. This number has been increasing sharply since 2009, recording the use of about 8.4 billion IoT devices as of January 2017 [1–3].

IoT is used in many fields according to its potential. IoT based Environment management system is also an important field that needs to be discussed in a national level. The environment management system monitors and manages the surrounding environment at all times. Smart farm, being a subset of environment management type, monitors the environmental information such as surrounding temperature, humidity, and carbon dioxide concentration.

Most of the existing IoT devices are located at the end service, as devices in home appliances sector, so even if an incident related to security happens, they are less affected. For that reason, security is weak with focus on cost and convenience. On the other hand, the industrial area is located in the supplier's position, and scale is in a large national level. Therefore, the influence of security incident is very large. Because of the introduction of IoT in the originally closed industrial area, many security vulnerabilities have been created and must meet the existing IoT and other conditions such as strong security, minimized control to humans, and strong resistance to the environment [4].

But despite this scale, the issue of IoT's vulnerable security is getting worse day by day. As can be seen in Table 1 below, there is an increasing number of malware targeted at Internet of Things every year [5]. This is due to the poor performance of IoT devices that are concerned with low-cost policy and thus poor security. Therefore, it is urgent to solve the security problems of IoT that holds such high importance.

Cryptography is appropriate for enhancing the security of an IoT based environment management type. IoT should consider the resource consumption due to cryptography, and it is most reasonable to use lightweight cryptography to enhance security, as low performance IoT devices at the end service also have to apply cryptography. However, it is necessary to find out what kind of lightweight cryptography is the optimal lightweight cryptography. Therefore, by implementing services that meet the conditions such as strong security in the industrial field, minimization of human control, strong resistance to the environment, and constant monitoring, CPU usage and processing time when each lightweight cryptography with security vulnerability is applied, helps to yield an optimal lightweight cryptography [6].

The composition of this paper is as follows. Section 2 introduces related research. In Sect. 3, we describe about IoT based devices of environment management type. Section 4 introduces test results of the cryptography implemented in the device. Finally, Sect. 5 discusses the conclusion of this paper.

Table 2 Implementation environment of environmental monitoring service

Type	Usage list
Hardware	Raspberry Pi 3 Arduino UNO DHT11 temperature and humidity sensor
Software	Raspbian Maria DB Python3 Idle Arduino Sketch Apache
Language	Python3 C++ SQL

2 Related Research

Developed countries like America, China, and Japan have already designed IoT technology, which is closely related to living environment such as intelligent power grid, as a major strategic industrial sector in the country. Various research and development in related academia are also being carried out [7]. Due to the performance limitations of IoT, research on applying lightweight encryption has been being pursued actively [8]. Based on a number of studies on single lightweight cryptography, international standardization [9], comparison and analysis of the processing time of encryption and decryption of several lightweight cryptography [10], a new cryptography for low powered devices have been studied [11]. Conversely, research has also been conducted to implement hardware tailored to specific lightweight cryptography. Development of small sized hardware for LEA [12], and hardware that is efficiently designed for LEA and CLEFIA to increase processing speed by 20–30% [13, 14] is being done. However, as in this paper, research on lightweight cryptography for certain types of services such as environment management type is hard to find.

3 Environmental Monitoring Service

In this paper, in order to derive a lightweight cryptography algorithm specialized for environment management IoT service, we implemented an Environmental Monitoring Service to monitor temperature and humidity continuously and to control temperature. The implementation environment is shown in Table 2 below, and overview is shown in Fig. 1.

DHT11 temperature and humidity sensor is a sensor that measures near temperature and humidity. Arduino Uno is a single-board microcontroller that connects to the sensor and receives temperature and humidity information. Arduino Uno is controlled by an Arduino sketch that uses a C++ programming language. Raspberry

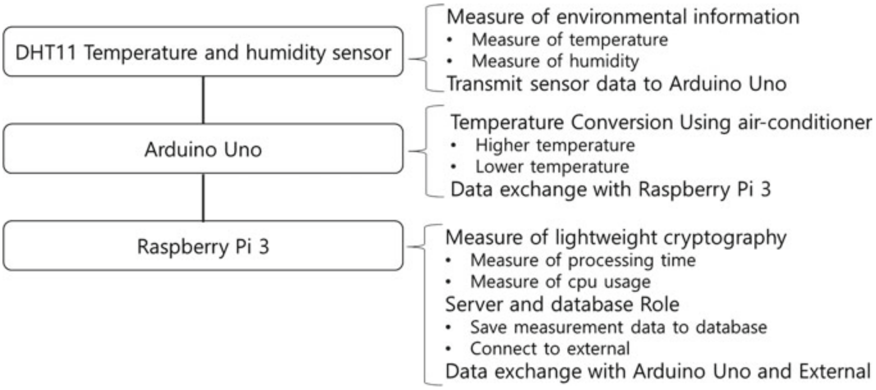

DHT11 Temperature and humidity sensor
Arduino Uno
Raspberry Pi 3
Measure of environmental information
• Measure of temperature
• Measure of humidity
Transmit sensor data to Arduino Uno
Temperature Conversion Using air-conditioner
• Higher temperature
• Lower temperature
Data exchange with Raspberry Pi 3
Measure of lightweight cryptography
• Measure of processing time
• Measure of cpu usage
Server and database Role
• Save measurement data to database
• Connect to external
Data exchange with Arduino Uno and External

Fig. 1 Overview of environmental monitoring service

Pi 3 uses Raspbian, a Devian-based OS, and connects to Arduino Uno. Raspberry Pi collects processing time, CPU usage, and other information in a database to measure performance of several lightweight ciphers. The higher the CPU usage, the higher the power consumption, which will affect energy conservation and battery life. Processing time is important for services that require fast response, such as security. Raspberry Pi also acts as a server role when connected to the external environment. It is handled using Python 3 Idle, Apache, MariaDB.

4 Test Result

Table 3 below summarizes the statistics of the results of testing Environmental Monitoring Service. In this paper, only lightweight ciphers are selected and tested in consideration of low average performance of IoT devices. Raspberry Pi 3 requires up to 25% CPU usage as shown in Table 3, even though DMIPS is 20 times more than the average 50–100 DMIPS required by the sensor [15–17]. It is essential to use lightweight encryption. Commonly used form of lightweight cryptography is lightweight block cryptography. Lightweight block cryptography is characterized as that the length of the key is proportional to security. Three of the encryptions with three kinds of bit size, i.e. 128, 192, and 256 bit are selected for length comparison. Selected AES (Advanced Encryption Standard), LEA (Lightweight Encryption Algorithm), and CLEFIA are commonly used lightweight cryptography. Also, we use CBC (Cipher Block Chaining), which is a common block cipher mode of operation.

Table 3 shows statistics for processing 10,000 blocks of 128 bit size. Even though 10,000 blocks were processed, the CPU usage difference was less than 1%, and there was no significant difference in the comparison of the key length. Therefore, it is recommended that the environment management based IoT devices have high

Table 3 Test result of environmental monitoring service

	Key length (bit)	Processing time (s)	CPU usage (%)
No cipher	0	0.169	1.247
AES	128	0.25	5.76
	192	0.251	5.889
	256	0.26	6.023
LEA	128	6.14	25.352
	192	7.029	25.477
	256	7.925	25.497
CLEFIA	128	0.713	15.678
	192	0.725	16.241
	256	0.729	16.375

Table 4 Summary of environmental monitoring service

	Best	Intermediate	Worst
Security	LEA	CLEFIA	AES
Processing delay	AES	CLEFIA	LEA
Power consumption	AES	CLEFIA	LEA

security even in higher bit cases, unless it is a special case such as very large amount of data. The focus should also be on the type of cryptography, not just the key length.

AES is suitable for ultra-light devices with short processing time and lower CPU usage. CLEFIA showed relatively low processing time and high CPU usage. LEA has long processing time and high CPU usage, but identification of any sort of external attacks have not been recorded in this cryptography. AES and CLEFIA have weaknesses in security and relatively LEA is more secure. Table 4 summarizes the following.

Environment management type devices are often powered in a fixed state and are continuously being measured, and the power consumption is also small at all times because the measurements do not require high performance. However, when it comes to consumption of solar energy or portable battery, the amount of energy used needs to be considered. This is because the rate of power consumption is directly proportional to the CPU usage.

In the case of security, we can take an example of smart farm. If a smart farm gets hacked, the crops can be damaged by disturbing the temperature to wrong value. So, it is necessary to have a high level of security, even if the security is likely to cause performance issues. Block ciphers are more secure as the key length is longer. And the fewer the known vulnerabilities, the better.

When it comes to processing delay of the system, even if the delay time is about 10 s, the operation speed is not affected because it does not affect the environment management system. Also, if you process 10,000 blocks at a time like this test, consider the amount of monitoring information sent by environment management system.

Therefore, LEA can be considered an optimized encryption for the environment management system with high security. Although CLEFIA has a short processing time, it has relatively high CPU usage due to security and AES is good to use if power management is extremely difficult, but it is hard to say that it is specialized for use in environment management system.

5 Conclusion

IoT is being used in various fields and its momentum is getting higher over time. However, the security is weaker than the rising momentum, causing many concerns. In this paper, we implemented the Environmental Monitoring Service to enhance the security of environment management type IoT devices related to IIoT(Industrial IoT) which is large enough to be used in a national level. And we measured the CPU usage and processing time by applying AES, LEA, and CLEFIA lightweight cryptography, and found that the lightweight cryptography suitable for environment management type is LEA.

In the future, we will try to derive lightweight cryptography that is specific to other types of IoT devices that are not based on environment management.

Acknowledgements This research was financially supported by Changwon National University in 2017–2018.

References

1. Crisp A (2015 Feb 17) Internet of things: prime time for satellite? http://www.nsr.com/news-resources/the-bottom-line/internet-of-things-prime-time-for-satellite/
2. Egham UK (2017 Feb 7) Gartner says 8.4 billion connected “things” will be in use in 2017, Up 31 percent from 2016. https://www.gartner.com/newsroom/id/3598917
3. Harapnuik D (2012 Oct 18) iPhone users driving nearly half of smartphone Web traffic. www.harapnuik.org/?p=3167
4. Chang HS, Jin KH, Shon T (2015.10) A study on cyber security issues in industrial IoT Environment. Rev KIISC 25.5:12–17
5. Kuskov V, Kuzin M, Shmelev Y, Makrushin D, Grachev I (2017 Jun 19) Honeypots and the internet of things. https://securelist.com/honeypots-and-the-internet-of-things/78751/
6. Ko YS Study of policies of major countries on internet of things and
7. Market Forecast (2014.12) Int Commer Inf Rev 16.5:27–47
8. Kim S-H, Jeong J-M, Hwang M-T, Kang C-S (2017.9) Development of an IoT-Based atmospheric environment measurement and analysis system. J Korean Inst Commun and Inf Sci 42.9:1750–1764

9. Kim M, Kim S, Kwon T (2016.2) Lightweight cryptographic technology trend for IoT communication environment. J Korean Inst Commun Sci 33.3:80–86
10. Jung Y-H, Song J (2015.8) ISO/IEC JTC 1/SC 27 WG2 International standardization trend of lightweight cryptography technology. Rev KIISC, 25.4:11–17
11. Seo H, Kim H (2015.4) Implementation of lightweight cryptographic algorithm for internet of things. Rev KIISC 25.2:12–19
12. Kim J-H (2017.2) A new type of lightweight stream encryption algorithm motif for applying low capacity messaging data encryption for IoT/QR/electronic tags. J Korea Inst Inf Electron Commun Technol 10.1:46–56
13. Sung M-J, Shin K-W (2015.4) A small-area hardware design of 128-bit lightweight encryption algorithm LEA. J Korea Inst Inf Commun Eng 19.4:888–894
14. Mi-Ji Sung, Kyung-Wook Shin (2015.7) An efficient hardware implementation of lightweight block cipher LEA-128/192/256 for IoT security applications. J Korea Inst Inf Commun Eng 19.7:1608–1616
15. Bae G-C, Shin K-W (2016.2) An efficient hardware implementation of lightweight block cipher algorithm CLEFIA for IoT security applications. J Korea Inst Inf Commun Eng 20.2:351–358
16. Voica A (2016 Jun 21) A guide to internet of things (IoT) processors. https://www.mips.com/blog/a-guide-to-iot-processors/
17. Longbottom R (2017 May) Roy Longbottom's Raspberry Pi, Pi 2 and Pi 3 Benchmarks. http://www.roylongbottom.org.uk/Raspberry%20Pi%20Benchmarks.htm

Improving an Evolutionary Approach to Sudoku Puzzles by Intermediate Optimization of the Population

Matthias Becker and Sinan Balci

Abstract In this work we improve previous approaches based on genetic algorithms (GA) to solve sudoku puzzles. Those approaches use random swap mutations and filtered mutations, where both operations result in relatively slow convergence, the latter suffering a bit less. We suggest to improve GA based approaches by an intermediate local optimization step of the population. Compared to the previous approaches our approach is superior in terms of convergence rate, success rate and speed. As consequence we find the optimum with one population member and within one generation in a few milliseconds instead of nearly one minute.

Keywords Sudoku · Genetic algorithms · Criticism

1 Introduction

Since a number of years, Sudoku puzzle became a popular spare time activity, also in Europe. The research community also adopted the problem to design and solve such puzzles, which are NP-complete in their complexity [13]. For such problems heuristic algorithms are a viable way of finding a solution, when a complete evaluation of the search space is not possible. Several works evaluated that possibility for Sudoku puzzles, among them also those using Genetic Algorithms (GA) [7, 8], presented at the Congress on Evolutionary Computation (CEC) in the past. Other works tackle the problem using SAT Techniques [6], linear integer programming [2] or enumeration approaches [4].

We particularly refer to the results of the article [14] presented at CEC 2015. The basic result of this work is an improvement of the GA based approach from [8] where the geometric crossover methodology is introduced, so that the improved approach using filtered mutations reveals a higher success rate of finding the solution (all except one test case out of nine versus five out of nine previously). Both approaches need

M. Becker (✉) · S. Balci
University Hannover FG Simulation, Welfengarten 1, 30167 Hannover, Germany
e-mail: xmb@sim.uni-hannover.de

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_38

roughly 50 seconds runtime. Additionally the influence of the tournament size on the success rate is tested using a more difficult Sudoku [5]. It shows that a tournament size of three is too few, and for sizes four to seven, the success rate ranges between 40 to 50%. The benchmark Sudokus are taken from [5, 9]. In our work, we seek further improvement of the GA based approach to solving sudokus and second, discuss whether GA is a suitable algorithm for that problem.

2 Our Approach

It has been shown in many works that the success and performance of GA can be improved to a great extend by combining global and local search steps [11]. The speed of the convergence is improved significantly if the current population of the GA is optimized with local search methods or case-specific knowledge in each step of the GA. That means in each step of the GA, the usual mutation, crossover and selection operators are applied and a new population is build up. The resulting population is then optimized using local search algorithms.

As shown in [11, 12], the combination of local and global search algorithms can significantly boost the performance of the search procedure. The combination can consist of GA and any other local search algorithm, such as Taboo Search, Simulated Annealing, Hill Climbing or other. The important aspect of the combination of local and global search algorithm is that on the one hand, the global search algorithm potentially covers a wide part of the search space by a considerable part of randomness and 'jumping' in the search space.

By this, it is avoided to get stuck in local optima, the randomness and the ability to 'jump' guarantees flexibility. On the other hand, too much 'jumping' decreases the speed of convergence towards the optimum. Even if the search space is structured very conveniently, e.g. being monotonic towards one global optimum, the GA with its global search characteristic will need much more steps towards the optimum compared to a greedy algorithm such as Hill Climbing. Especially when applied to problems with very high complexity and many constraints (as shown e.g. in [3]), probabilistic search might fail to find even one single neighbor that would fit to the constraints. The mixing of algorithms should in the best case combine the advantages of both sides, avoiding the negative properties of each algorithm.

Thus in our approach, the GA fulfills a global search and avoids/escapes local optima and the local search refines the individual members of the population, thus avoiding the crossover of sub-optimal population members and the slow convergence rates of GA when near the global optimum. In that case, the global search characteristics of the GA often might *leave* the area where the global optimum is located.

Algorithm 1 Local Optimization of the Population in one Iteration

```
Locally optimize p:
repeat
    Gap can be filled by logic deduction
until No more gap found or Time limit reached
if Time left then
    Search neigborhood recursively
else
    Stop local optimization and proceed with next iteration of GA
end if
```

2.1 Algorithm

As described above, our algorithm works in general as a GA, but is improved by an intermediate local optimization of the population members in each step of the GA. As local search, we use a greedy algorithm such as hill-climbing, and additionally we use case-specific knowledge to enhance the members of the population. The Sudoku-specific knowledge here is the fact, that in a given population member, often several unknowns can be deduced non-ambiguously using the rules of the game. E.g. if in on row only one number is missing. Using a time-limited recursive local search algorithm that is able to fill the gaps that can be logically deduced, those gaps are filled much more efficiently than by the guided random search scheme of the GA.

Since the performance of the previous work to beat is in the order of one minute, the time scheduled for the local optimization should be set several orders of magnitude lower than that, i.e. in the order of milliseconds. The sketch of the algorithm can be found in algorithm 2.

Algorithm 2 GA enhanced by local optimization

```
Initialize Population
repeat
    Carry out GA operations: crossover, mutation, selection
    for all p in Population P do
        repeat
            Locally optimize p (Algorithm 1)
        until Time limit reached or Local optimum found
    end for
until Max. number of iterations reached or Quality of solution sufficient
```

The test-cases are the same as in [14], taken from [9] except two taken from [5] and are shown in Table 1.

Table 1 Benchmarks

Difficulty	Givens
Easy1	38
Easy2	34
Medium1	30
Medium2	29
Difficult1	28
Difficult2	24
Super difficult1	24
Super difficult2 (AI escargot) [5]	23
Super difficult3	22
Hardest [5]	21

3 Experiments

First we had to do some initial experiments in a pre-run in order find the adequate amount of calculation time allowed for the local search, since if the time is to short, the benefit will be not significant, if too much time is spent in the local search, then the local search might lower the performance of the GA significantly. Thus for the GA, we adopted the parameters from the previous work, and in order to find the optimal balance of the time used by the GA calculations and the time devoted to the recursive local search, we varied the time for the local search performed on the complete population of one generation of the GA from one millisecond to 1000 ms. The outcome can be observed in Fig. 1. On the x-axis, the time in milliseconds allowed for local search in each iteration of the GA is given, on a logarithmic scale. On the y-axis it is noted, for how many classes of difficulty a solution can be found faster than using pure GA. (All test cases are rated due to the difficulty and ordered in five classes as can be seen in Table 3).

We conclude from these pre-runs, that a majority of sudokus up to category 'super difficult2' can be solved much faster when the time for local search is below 15 ms, and only for the category 'hardest', more time (250–500 ms) should be given to the local optimization for an overall speedup of finding the solution.

3.1 Final Experiments

Outgoing from the pre-runs the parameters for the final experiment for the comparison to the performance to previous approaches has been determined. As explained above, the parameter concerning the GA had been used as in previous works in order to have the results in this work comparable to the previous works as baseline. For our

Fig. 1 Efficiency of the time assigned to local search

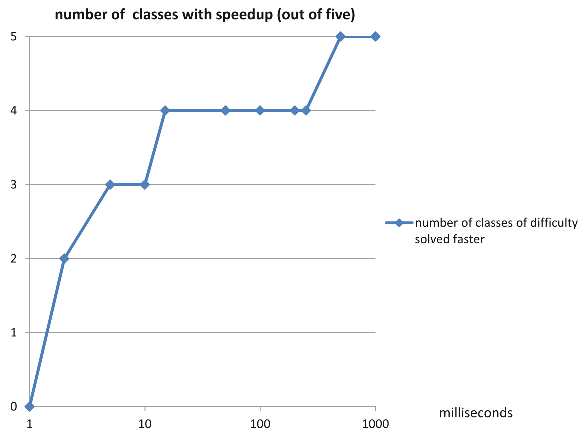


Table 2 Success rates

Difficulty	Success rate [14]	Success rate (our approach)
Easy1	100	100
Easy2	100	100
Medium1	100	100
Medium2	100	100
Difficult1	100	100
Difficult2	100	100
Super difficult1	100	100
Super difficult2 [5]	100	100
Super difficult3	96	100
Hardest [5]	38	100

extension of the algorithm, that is the introduction of problem specific knowledge and local optimization onto the population of the GA in each iteration, we used the findings of the pre-runs in order to balance the local and global search characteristics of the combined algorithm and eventually set the maximum time given to the additional effort of the direct optimization of the population to 300 ms.

However it is possible that the combined algorithm succeeds in less than 300 ms: If the local search is more efficient than the global search then it is possible that the algorithm will halt in less than 300 ms: Then obviously the sudoku problem is more amenable to a local search algorithm and can be solved efficiently using that characteristic algorithm.

In this way, our combined algorithm in a way ‘decides’ how much weight should be laid on the local and the global search depending on the problem presented and chooses the right balance between both parts.

Table 3 Speed of calculation

Difficulty	Duration [14] (ms)	Duration (our approach) (ms)
Easy	48040	1.5
Medium	49240	2
Hard	48830	3.7
Super difficult2 [5]	49300	14
Hardest [5]	48760	257

4 Results

In this section we compare the results of our approach with the previous success rates and performance. In several runs we finally could lower the population size and number of iterations to very low values for each benchmark. The calculations have been carried out on a normal desktop computer comparable to the previous works and were not parallelized.

The results concerning the success rates can be found in Table 2.

In the previous work [14], no standard deviation is given. For the Sudoku 'Hardest' we give the average of the results that have been conducted for evaluation of a good value for the tournament size from the original publication. The values there ranged between 6 and 51%, the average success rate thus is 38%.

As can be seen, our approach always finds the solution, thus has a success rate of always 100%, that is why no standard deviation is given.

The results concerning the runtime can be found in Table 3. Since in [14] the runtime results are given for each class of difficult, we arrange our table accordingly.

It can clearly be seen that our approach using the intermediate optimization of the populations found in the GA is superior concerning the speed of the algorithm to the previous approach by several orders of magnitude.

5 Conclusion and Discussion

In this work we present the improvement of an approach based on Genetic Algorithms. The improvement consists out of the intermediate optimization of each population of the GA by greedy local search and problem specific improvement of each member of the population. It turns out that this combined approach is superior to the previous in terms of the search success as well as in performance. Especially the performance is better in orders of magnitude. This can be explained by the fact, that the Sudoku puzzle has many constraints which are not easy to fulfill by the search of the GA which has a large component of randomness. This more global search characteristic might work well with problems that do not impose many constraints on the solution. For the Sudoku puzzle the application of pure GA results in too large runtime and in a search success not always reaching 100%.

The conclusion should be that heuristic search procedures are not a well suited means for tackling the standard Sudoku puzzle. Seeking help from naturally inspired algorithms is a good idea in many cases, however it should not be done without critical considerations whether the algorithm is suitable at all, or whether other approaches might promise more success. See also for critical considerations about maze solving of pseudo intelligent natural paradigms. See also [1, 10] where it is argued very convincingly that the metaphor thinking should not be overstretched.

Of course it is a good idea to seek for inspiration for new algorithms and new paradigms, however the usefulness and the efficiency should be honestly evaluated. For application of pure GA we have shown that this algorithm is not a good choice, it should at least be helped by local search and problem specific knowledge.

References

1. Barras C (2010) What a maze-solving oil drop can tell us about intelligence. New Sci 205(2744):8–9
2. Bartlett A, Chartier TP, Langville AN, Rankin TD (2008) An integer programming model for the Sudoku problem. J Online Math Appl **8** (2008)
3. Becker M (2006) Genetic algorithms for noise reduction in tire design. In: 2006 IEEE international conference on systems, man and cybernetics, vol 6, pp 5304–5308, Oct 2006. https://doi.org/10.1109/ICSMC.2006.385151
4. Felgenhauer B, Jarvis F (2005) Enumerating possible Sudoku grids. http://www.afjarvis.staff.shef.ac.uk/sudoku/sudoku.pdf
5. Inkala A (2007) AI Escargot—the most difficult sudoku puzzle. Lulu Publisher, Finland
6. Lynce I, Ouaknine J (2006) Sudoku as a sat problem. In: ISAIM
7. Manter T, Koljonen J (2007) Solving, rating and generating Sudoku puzzles with GA. In: IEEE congress on evolutionary computation, 2007. CEC 2007. IEEE, pp 1382–1389 (2007)
8. Moraglio A, Togelius J, Lucas S (2006) Product geometric crossover for the Sudoku puzzle. In: IEEE congress on evolutionary computation, 2006. CEC 2006. IEEE, pp 470–476 (2006)
9. Sato Y, Inoue H (2010) Solving Sudoku with genetic operations that preserve building blocks. In: 2010 IEEE symposium on computational intelligence and games (CIG). IEEE, pp 23–29 (2010)
10. Sörensen K (2013) Metaheuristics—the metaphor exposed. Int Trans Oper Res (2013)
11. Syrjakow M, Szczerbicka H (1999) Efficient parameter optimization based on combination of direct global and local search methods. In: Evolutionary algorithms. Springer, pp 227–249
12. Syrjakow M, Szczerbicka H, Becker M (1998) Genetic algorithms: a tool for modelling, simulation, and optimization of complex systems. Cybern Syst 29(7):639–659
13. Takayuki Y, Takahiro S (2003) Complexity and completeness of finding another solution and its application to puzzles. IEICE Trans Fundam Electron Commun Comput Sci 86(5):1052–1060
14. Wang Z, Yasuda T, Ohkura K (2015) An evolutionary approach to Sudoku puzzles with filtered mutations. In: 2015 IEEE congress on evolutionary computation (CEC). IEEE, pp 1732–1737 (2015)

Saturated Flow Method for Traffic Optimization in Ostrava City

Martin Kotyrba, Eva Volna and Jakub Gaj

Abstract This article discusses the possibility of verifying whether the formalism of Petri nets suitable for the optimization of transport systems. This work discusses the optimization of automotive traffic using the saturation flow method. The main aim of the article is to propose and validate a methodology for optimizing transport systems with saturation flow methods. There are discussed several optimization methods that are used for this purpose. The results of this search section were summarized and then a suitable model and an optimization method were chosen according to established criteria. Based on them we have proposed methodology that was verified pursuant to the application of the chosen procedures. One of Ostrava's most interrupted crossroads was chosen as a model for optimization and simulation. Results of the experimental study are summarized in the conclusion. From the results of the experimental study, it can be stated that the static signaling model has been optimized in terms of time.

Keywords Saturated flow · Traffic optimization · Petri nets

1 Introduction

In this article, we have focused on the various steps of the proposed optimization methodology to ease traffic and enable smoother transit. It should be noted that most of the traffic junctions are controlled dynamically by means of sensors, but at the moment of any failure the operation is automatically switched to the static control

M. Kotyrba (✉) · E. Volna · J. Gaj
Department of Informatics and Computers, University of Ostrava, 30 Dubna 22,
70103 Ostrava, Czech Republic
e-mail: martin.kotyrba@osu.cz

E. Volna
e-mail: eva.volna@osu.cz

J. Gaj
e-mail: jakub.gaj@osu.cz

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_39

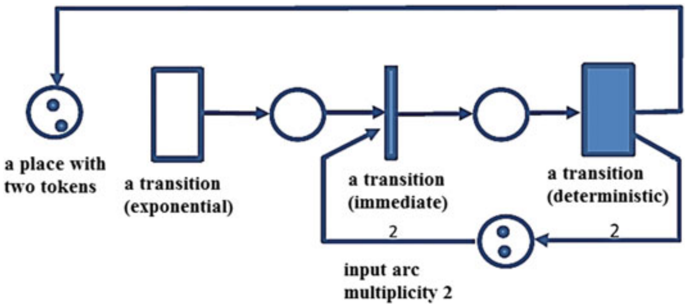


Fig. 1 Example of Petri Net Adapted from http://www.peterlongo.it/Italiano/Informatica/Petri/index.html

model, which has undergone the final optimization in this work. Using the saturation flow method, we set the cycle lengths and the free (green) signal depending on the saturation levels of the entrances in the individual phases. The basic calculation period for a proposal with this method is a 1-h time period.

2 Petri Nets

Petri nets, introduced by C. A. Petri in 1962 [1], provide an elegant and useful mathematical formalism for modeling concurrent systems and their behaviors. In many applications, however, modeling by itself is of limited practical use if one cannot analyze the modeled system. As a means of gaining a better understanding of the Petri net model, the decidability and computational complexity of typical automata theoretic problems concerning Petri nets have been extensively investigated in the literature in the past four decades. Petri nets are a promising tool for describing and studying systems that are characterized as being concurrent, asynchronous, distributed, parallel, nondeterministic, and/or stochastic. As a graphical tool, Petri nets can be used as a visual-communication aid similar to flow charts, block diagrams, and networks. In addition, tokens are used in these nets to simulate the dynamic and concurrent activities of systems. As a mathematical tool, it is possible to set up state equations, algebraic equations, and other mathematical models governing the behavior of systems.

An example of Petri Net is shown in Fig. 1. Definition [2]: Petri net is a net of the form PN = (N, M, W), which extends the elementary net so that:

- N = (P, T, F) is a net, where P and T are disjoint finite sets of places and transitions, respectively. $F \subset (P \times T) \cup (T \times P)$ is a set of arcs (or flow relations).
- M: $P \rightarrow Z$ is a place multiset, where Z is a countable set. M extends the concept of configuration and is commonly described with reference to Petri net diagrams as a marking.

- W: F → Z is an arc multiset, so that the count (or weight) for each arc is a measure of the arc multiplicity.

If a Petri net is equivalent to an elementary net, then Z can be the countable set {0,1} and those elements in P that map to 1 under M form a configuration. Similarly, if a Petri net is not an elementary net, then the multiset M can be interpreted as representing a non-singleton set of configurations. In this respect, M extends the concept of configuration for elementary nets to Petri nets [2].

3 Optimization Methods

In order to assess whether the monitored process works effectively, we need to determine the criteria by which we should optimize. We come to the question of what is the most appropriate criterion for assessing the effectiveness of the process. Most often we build on our own or acquired experience. The criterion for assessing effectiveness is the value of the effect flowing to its users. When searching for an optimal solution, we are limited to so-called restrictive conditions. The proposed solution that accepts the set limits is called an acceptable solution.

Optimization methods should be reliable in finding optimal solution and should be fast, which means they should be able to find optimum efficiently. In order to meet the first condition, we need to perform a so-called optimality test. We understand the optimality test as a rule we determine, based on its use, whether the solution that is currently being tested is optimal or not. Optimization methods can be divided into primary and dual. The basic difference is that the primary are moving entirely in the field of acceptable solutions, which is already the default solution and we gradually improve it, the dual methods can start within an inaccessible solution.

Another look at the division of optimization methods is the division to the exact and heuristic. With exact methods, we have the possibility to test whether the solution satisfies the optimality condition, we do not have this option for heuristic methods. Heuristic methods give us a solution that we don’t know about whether it is optimal. If there is a sufficiently powerful exact method, we always prefer it to any heuristic.

In general, two basic forms of light control are distinguished - static and dynamic. The basic difference between them is that, while static control, the signal plan does not adjust according to the specific situation, with dynamic control does. However, even junctions that are dynamically controlled have a static signaling plan in backup, which is immediately activated in the case of a dynamic control failure [3].

3.1 Time Consumption Method

The principle of this method is that the intensity of individual traffic directions changes by multiplying the coefficient of the limit factor. This coefficient takes into

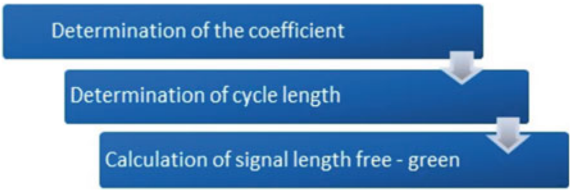

Determination of the coefficient
Determination of cycle length
Calculation of signal length free - green

Fig. 2 The optimization phase of time consumption method

account the effect on deceleration or acceleration of the vehicle over the intersection. This fictitious so-called computational load is introduced into the calculation of the cycle length and the individual green phases according to (1).

$$M = \frac{I.k}{n}[j.v./h] \tag{1}$$

where M is the computational load, I is the intensity of a particular traffic direction, k illustrates the resulting coefficient of the limit factor for the traffic direction and n is the number of shift lanes of this direction as shown in Fig. 2.

3.2 Linear Programming Method

There are a number of methods that have been developed for this purpose but the basic advantage of linear programming is that the design of the signal plan running objectively without significant interventions of the investigator, which is not met in other methods. The task is to decide on the moments of the beginnings and endings of greens for each stream so that:

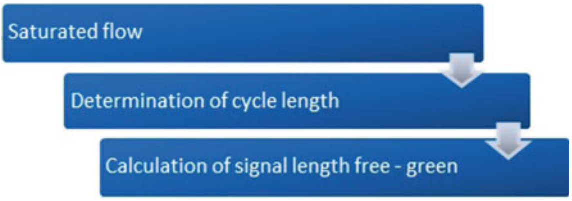
a. at the specified value of minimum proportional reserve, between the offered and the average required green time for the stream, minimized the cycle length,
b. at the specified cycle length, to maximize the value of minimal reserve (as defined in point a).

By selecting criterion (a) we achieve a solution in which more cycles occur during the selected time unit (individual streams will enter the intersection more often), by choosing criterion (b) we will achieve greater fairness in allocating the time of green to individual streams (depending on the average green time required), [4].

3.3 Saturated Flow Method

Using the saturation flow method, we set the cycle lengths and the free (green) signal depending on the saturation levels of the entrances in the individual phases. The basic calculation period for a proposal with this method is a 1-h time period [4].

Fig. 3 Phase of optimization method of saturated flow


Saturated flow
Determination of cycle length
Calculation of signal length free - green

Saturation flow is the highest number of vehicles that can travel through the stop-line profile per unit of time under ideal traffic conditions [u.v./hr]. The individual symbols indicate the following: u.v.—unit vehicles, hr.—hour. Phase of optimization method of saturated flow is shown in Fig. 3.

Saturated Flow.

Saturated flow depends primarily on:

- entrance width,
- longitudinal slope,
- radius of arc,
- portion of turning vehicles.

The basic saturated flow of the shift lane depends only on the width of the shift lane (2). For collective roads, bidirectional, 4 and multilane or 1-way, 2 and multilane with $v = 50 - 60$ km/h with good road surface [5].

$$S_{basic.(line)} = 2000\, j.v./h \tag{2}$$

Basic saddle entrance by condition that the entrance is formed by one shift lane (3):

$$S_{basic.} = S_{basic.(line)} \tag{3}$$

If the entrance is made of more lanes then (4):

$$S_{basic} = \sum S_{basic.(line)} \tag{4}$$

Saturated flow entry is determined from the basic saturated flow of entrance according to (5).

$$S = S_{basic}.k_{inclination}.k_{arc} \tag{5}$$

The coefficient $k_{inclination}$ is the coefficient of inclination (6), which expresses the influence of the inclination on the saturation flow.

$$k_{inclination} = 1{-}0,02 \quad a \tag{6}$$

This formula applies to climb to 10%. If the road is horizontal or decreasing, the value a is set to 0, in case that the entry increases by more than 10%, we set the value a to 10.

The coefficient k_{arc} expresses the effect of the radius of the arc at the branch and the share of the branching vehicles to the saturated flow according to (7). Applies in general for turning right and left [5].

$$k_{arc} = \frac{R}{R + 1,5.f} \tag{7}$$

R is the designation of the radius of the arc while turning in meters. Furthermore, f is the ratio of the departing vehicles of the total entrance intensity according to (8)

$$f = \frac{the\,intensity\,of\,turning\,vehicles[car/h]}{the\,total\,intensity\,of\,the\,entrance[car/h]} \tag{8}$$

If there is a separate turning lane, we set $f = 1$. If for left turn exists a common shift lane with straight direction or with right turn and at the same time the left turn is influenced by the counter direction (in the same phase, the left turning on camming vehicles must give priority to them), it expresses the effect of giving priority over turning vehicles to the left to reduce the capacity of entry specifying a fictitious arc radius (the input is uniformly $R = 1.5$ m).

In case the right turn (on a separate and common lane) is significantly influenced by the flow of parallel pedestrians, the effect is expressed to giving priority to pedestrians by the right turn on vehicles to reduce the entry capacity by entering the fictitious arc radius. This is entered according to the intensity of the pedestrians passing [5].

Determination of Cycle Length.
Determine the degree of saturation according to (9)

$$y = \frac{I}{S} \tag{9}$$

for all entrances with vehicular traffic. Other entrances are ignored. In each phase, we select the entry that has the highest saturation level, the highest y, therefore the critical entry at the stage. The sum of these entrances we get the overall saturation degree according to (10).

$$Y = \sum_{i=1}^{n} \max y_i \tag{10}$$

In this formula i is the i-th phase and n is the number of phases.

In the next step, we have to determine the loss time for each phase. It is based on the assumption that at each phase we have a productive or effective green labeled z'. Effective green z' is the time during which the vehicle passing through the stop line in saturation flow. The z' value is obtained by summing the length of the green and the yellow part at the entrance of the last vehicle and by subtracting the time loss resulting from the vehicle startup response according to (11).

$$z' = z + 2 - 1 = z + 1[s] \tag{11}$$

The loss time for each phase l is the time between the end of the effective green at this phase and the start of the effective green in the next phase, i.e. the non-productive time during a changing of a phase according to (12).

$$l = t_m - \left(z' - z\right) = t_m - 1\,[s] \tag{12}$$

The loss time for each phase is determined by the actual split time t_m between the critical entrances in ending and next phase according to the structural signal plan. Sum of all loss times of each phase we get the total loss time per cycle according to (13).

$$L = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} t_{mi} - n\,[s] \tag{13}$$

In this formula, i is the i-th phase, n is the number of phases, l_i is a loss time for the i–th phase, t_{mi} is the time between the critical entries at the i-th and subsequent phases.

The optimal C_{opt} cycle is the cycle according to (14), in which the total delay of randomly arriving vehicles under the given conditions is minimal. It depends on the phase scheme, split times, and intensity of the traffic [5].

$$C_{opt} = \frac{1,5.L + 5}{1 - Y} \tag{14}$$

The optimum cycle is the basis for the design of a real cycle, which must be modified with respect to the trams and the long clearing times of the pedestrians at the crossings. Based on the calculated optimal cycle, it is possible to design a real cycle according to (15) in the range:

$$0.75 \cdot C_{opt} < C < 1.5 \cdot C_{opt} \tag{15}$$

Within this range applies that the real cycle approaches the optimal cycle and time loss of randomly arriving vehicles does not change much. The real cycle cannot be shorter than the structural cycle. Lengths of real cycles for control should not be greater than 100 s. Exceptionally, the cycle time can be up to 120 s. With fixed lengths of cycles greater than 120 s, excessively long stagnation increases. If this is appropriate, a higher maximum cycle length may be used for dynamic multi-phase control with variable cycle lengths.

Calculation of the Lengths of Free—Green Signals
In the first step, we determine the lengths of greens for critical entrances in the individual phases according to (16):

$$z = \frac{y.(C - L)}{Y} - l\,[s] \tag{16}$$

These greens of critical entries in the individual phases determines the optimal lengths of each phase of the signaling plan [5].

4 Proposal of Optimization Method

At the moment when we have a real traffic situation model, we move to the optimization method selection phase. We choose the criterion we are trying to optimize. Based on the selected criteria we use methods that optimize this criterion. In my case, I count the number of cars per hour in the traffic peak on given intersection, i.e. between 15:00–16:00 on a working day and trying to optimize the situation so that the given number of vehicles should pass in a shorter time period than 1 h. Therefore, the saturation flow method was chosen and based on this method, the cycle length and the free (green) signal value were calculated. The preliminary studies of the proposed optimization method were published in [6, 7].

4.1 Applying the Optimization Method to a Selected Model

After selecting optimization method, we lead to the application of optimization method on selected model. The first step is to set the phase diagram. We determine which stream will go in which phase. In the selected model, the five phases were determined, see Fig. 4.

Next, we determine the proportion of turning vehicles in directions that can go straight or turning. In the selected model, we only have two directions like this, in other cases, there is always a separate strip for turning. Then we determine the radius of the arcs for turning and the split time table. Now we begin with the calculation of slope coefficients k_{inc} according to formula (6) and arcs k_{arc} according to formula (7). After calculating the coefficient, we obtain the saturated flow S of the given entry according to the formula (5), and the degree of saturation y of the individual entrances according to formula (9), see Table 1.

We divide the individual streams into the phases and assign saturation (Table 2). At each stage, we select a higher degree of saturation, which marks the critical entry at that stage. The sum of the saturation sessions of the critical entrances gives the

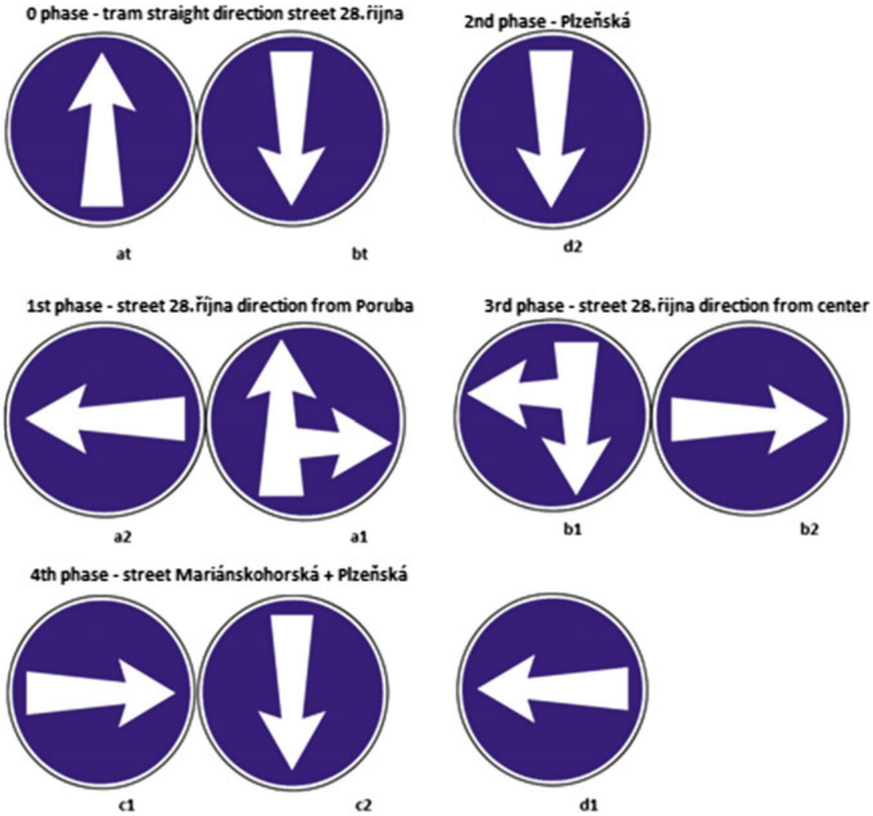

0 phase - tram straight direction street 28.října
at
bt
2nd phase - Plzeňská
d2
1st phase - street 28.října direction from Poruba
a2
a1
3rd phase - street 28.října direction from center
b1
b2
4th phase - street Mariánskohorská + Plzeňská
c1
c2
d1

Fig. 4 Phase diagram

overall degree of saturation according to formula (10), $Y = 0.778$. Now we obtain the optimum cycle length according to formula (14), $C_{opt} = 130.61$ and then the real cycle length ($0.75 \cdot C_{opt} < C = 240 < 1.5 \cdot C_{opt}$) according to formula (15) and the free green signal length for the critical passes at the given stage (Table 3) according to formula (16) and therefore also for the entire phase in which the current is present. The last step is the construction of the signaling plan, see Fig. 5.

5 Conclusion

The use of the saturation flow method is therefore appropriate for determining the cycle length and the determination of green time for each phase and a creation of the signaling plan, so setting the model and starting the simulation led to the verification of the model. A counter has been added to the model, which is set to 3600 steps,

Table 1 Saturated flow

Traffic stream	α (%)	k_{inc}	f (%)	R (m)	k_{arc}	l (j.v./h)	S (j.v./h)	y
a1	0	1	0.03	29.5	1.00	735	3994	0.184
a2	0	1	1	35	0.96	766	3836	0.200
at	0	1	0	1	1.00	29	2000	0.015
b1	0	1	0.40	16	0.96	755	3857	0.196
b2	0	1	1	50	0.97	48	1942	0.025
bt	0	1	0	1	1.00	30	2000	0.015
c1	0	1	0	1	1.00	539	4000	0.135
c2	0	1	1	23	0.94	1048	3755	0.279
d1	0	1	0	1	1.00	399	2000	0.200
d2	0	1	1	45.5	0.97	343	3879	0.089

Table 2 Determination of critical entrances

Traffic stream	Phase	y	
a1	1	0.184	
a2	1	0.200	
at	0	0.015	
b1	3	0.196	
b2	3	0.025	
bt	0	0.015	
c1	4	0.135	
c2	4	0.279	
d1	4	0.200	
d2	2	0.089	

Table 3 Determination of green time in critical phase inlets

Traffic stream	Phase	y	z	Opt.
at	0	0.015	3.18	5
a2	1	0.200	56.23	57
b1	3	0.196	55.39	55
c2	4	0.279	79.39	80
d1	2	0.089	24.52	25

simulating 3600 s or 1 h, as one tick in the model simulates 1 s. To stop the simulation, all traffic streams for cars have been emptied or when the time has elapsed for 1 h. After the simulation of the given model with setup time off signals—green according to the method of saturated flow, all vehicles passed through the intersection in 51 min 4 s, which means saving time 8 min and 56 s per hour during rush hour. Therefore, we can state that the static signaling model has been optimized in terms of time.

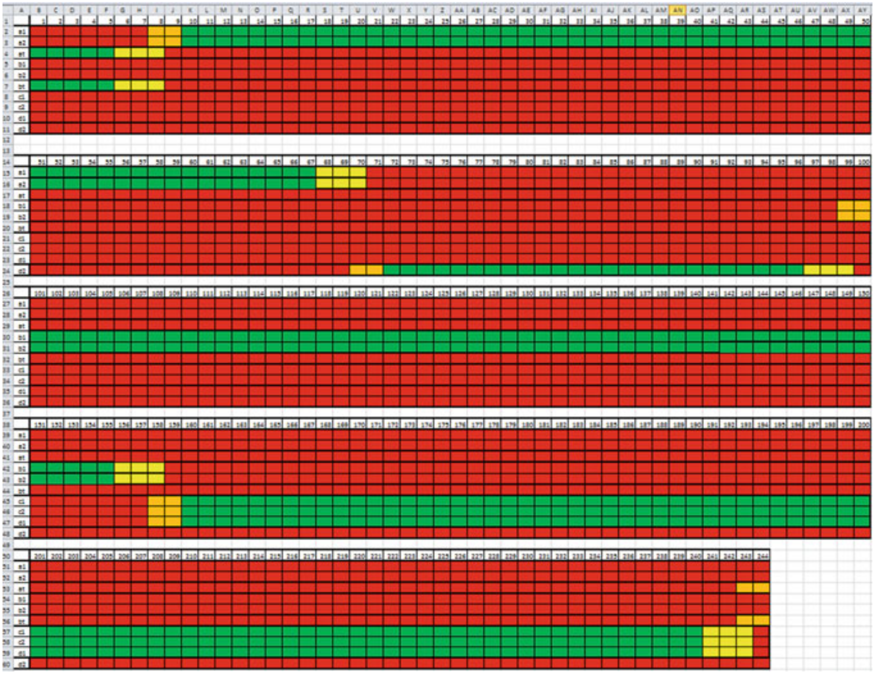


Fig. 5 Signal plan

Acknowledgements The research described here has been financially supported by the University of Ostrava grant SGS/PrF/2018.

References

1. Petri CA (1962) Communication with automata, supplement 1 to technical report RADC-TR-65-337, NY, 1965. Translation by CF Greene of Kommunikation mit Automaten. Diss. PhD Dissertation, University of Bonn
2. Rozenburg G, Engelfriet J (1998) Elementary net systems. In: Lectures on petri nets I: basic models–advances in petri nets. LNCS, vol 1491. Springer, pp 12–121
3. Devillers R, Antti V (2015) Application and theory of petri nets and concurrency. Springer International Publishing
4. Ng KM, Reaz MBI, Ali MAM (2013) A review on the applications of Petri nets in modeling, analysis, and control of urban traffic. IEEE Trans Intell Transp Syst 14(2):858–870
5. Křivda V (2004) Design of a light signaling device—calculations of fixed signaling plans (in Czech, VSB-TU) in Ostrava, Ostrava (2004) (Citation 19. 2. 2015). http://kds.vsb.cz/ord/ppt/ORD-SSZ-navrh.ppt
6. Gaj J, Kotyrba M, Volna E (2016) Possibilities of control and optimization of traffic at crossroads using petri nets. In: Information Science and applications (ICISA) 2016. Springer, Singapore, pp 21–29
7. Kotyrba M, Gaj J, Tvarůžka M (2017) The methodology for modeling queuing systems using Petri nets. In: AIP Conference Proceedings, vol 1863, no 1. AIP Publishing

Adaptive Control of EV3 Robot Using Mobile Devices and Fuzzy Logic

Jan Konvicka, Martin Kotyrba, Eva Volna, Hashim Habiballa and Vladimir Bradac

Abstract The main aim of this article is to show what is possible to use mobile devices with verification of adaptivity. We focused on creating applications for the control of robots Lego Mindstorms EV3 verification adaptivity, which uses fuzzy approach. We have used classical fuzzy rules of if-then type. The antecedent contains the measured values from infrared sensors and the consequent contains action response of individual engines of the robot. With this application, it will be able to control the robot via Wi-Fi while, there will be a possibility to bring the robot mode that will move through the maze without bumping into some of the walls. An integral part of the work is also a theoretical basis for adaptive robot control and autonomy. Part of this article describe of creating of applications, their comparison to other existing applications and experiments with the resulting application. The application perfectly functioned on the created experimental environments, including the adaptive mode.

Keywords Mobile devices · Mindstorm EV3 · Adaptivity · Fuzzy logic

J. Konvicka · M. Kotyrba (✉) · E. Volna · H. Habiballa · V. Bradac
Department of Informatics and Computers, University of Ostrava, 30 Dubna 22,
70103 Ostrava, Czech Republic
e-mail: martin.kotyrba@osu.cz

J. Konvicka
e-mail: jan.konvicka@osu.cz

E. Volna
e-mail: eva.volna@osu.cz

H. Habiballa
e-mail: hashim.habiballa@osu.cz

V. Bradac
e-mail: vladimir.bradac@osu.cz

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_40

1 Theoretical Background

The objective of this paper was to create applications for mobile devices which will enable to control robot Lego MindsStroms with subsequent verification of adaptivity on a selected problem. It was necessary to propose and equip the Lego MindStorms EV3 robot with suitable sensors, to create a set of fuzzy rules for four infrared sensors, to create an application to control it. Next steps consisted in implementing the fuzzy rules and verifying the functionality of the whole assembly on experiments.

1.1 Adaptive Control

Adaptive control is an active field in the design of control systems to deal with uncertainties. The key difference between adaptive controllers and linear controllers is the adaptive controller's ability to adjust itself to handle unknown model uncertainties. Adaptive control is roughly divided into two categories: direct and indirect. Indirect methods estimate the parameters in the plant and further use the estimated model information to adjust the controller. Direct methods are ones wherein the estimated parameters are those directly used in the adaptive controller [1].

Adaptive control is the control method used by a controller which must adapt to a controlled system with parameters which vary, or are initially uncertain. For example, as an aircraft flies, its mass will slowly decrease as a result of fuel consumption; a control law is needed that adapts itself to such changing conditions. Adaptive control is different from robust control in that it does not need a priori information about the bounds on these uncertain or time-varying parameters; robust control guarantees that if the changes are within given bounds the control law need not be changed, while adaptive control is concerned with control law changing themselves.

Conventional controllers are designed to control dynamic systems where the parameters do not change or where the parameters do not vary excessively when working around an operating point. However, it is very common to find systems where its dynamics change nonlinearly in an instant. In these cases, the conventional controllers do not behave as intended they do in all situations. This is the reason why adaptive controllers are used. In order to get these results using STR, a similar scheme to that of Fig. 1 is used, where typical closed loop control is shown. A second loop is added to identify the system parameters and to calculate the adaptive controller parameters.

In order to get these results using STR, a similar scheme to that in Fig. 1 is used, where typical closed loop control is shown. A second loop is added to identify the system parameters and to calculate the adaptive controller parameters [1].

There are several drawbacks in the use of adaptive controllers, such as the tuning is more complicated than that of a classical PID and the fact that a method to adjust the regulator without actually knowing the system dynamics is needed [1].

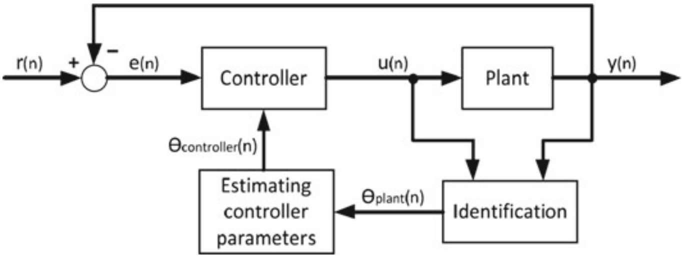


Fig. 1 General scheme overview of the adaptive control

1.2 Mobile Device with Android

A mobile device is basically any handheld computer. It is designed to be extremely portable, often fitting in the palm of your hand or in your pocket. Some mobile devices are more powerful, and they allow you to do many of the same things you can do with a desktop or laptop computer. These include tablet computers, e-readers, and smartphones. A smartphone is a powerful mobile phone that is designed to run a variety of applications in addition to providing phone service. Smartphones are basically small tablet computers, and they can be used for web browsing, watching videos, reading e-books, and playing games. Smartphones use touchscreens and operating systems similar to those used by tablet computers. Many of them use a virtual keyboard, but others have a physical keyboard, which allows the entire screen to be used for display purposes. Internet access is an important feature of smartphones. Generally, you will need to purchase a 3G, 4G, or LTE data plan in addition to normal cell service. Smartphones can also connect to Wi-Fi when it is available; this allows you to use the Internet without using up your monthly data allotment. Because they are optimized for Internet use, tablet computers have built-in Wi-Fi. For a monthly fee, you can also purchase a 3G or 4G data plan, allowing you to access the Internet from almost anywhere [4].

Android is a mobile operating system (OS) currently developed by Google, based on the Linux kernel and designed primarily for touchscreen mobile devices such as smartphones and tablets. Android's user interface is mainly based on direct manipulation, using touch gestures that loosely correspond to real-world actions, such as swiping, tapping and pinching, to manipulate on-screen objects, along with a virtual keyboard for text input. Variants of Android are also used on notebooks, game consoles, digital cameras, and other electronics [5].

Android's source code is released by Google under open source licenses, although most Android devices ultimately ship with a combination of open source and proprietary software, including proprietary software required for accessing Google services. Android is popular with technology companies that require a ready-made, low-cost and customizable operating system for high-tech devices. Its open nature has encouraged a large community of developers and enthusiasts to use the open-source code as a foundation for community-driven projects, which add new features for advanced users or bring Android to devices originally shipped with other operating systems. At the same time, as Android has no centralised update system most Android devices fail to receive security updates: research in 2015 concluded that almost 90% of Android phones in use had known but unpatched security vulnerabilities due to lack of updates and support. The success of Android has made it a target for patent litigation as part of the so-called "smartphone wars" between technology companies.

For development of our application we used leJOS [2, 7]. It is a firmware replacement for Lego Mindstorms programmable bricks. It currently supports the LEGO RCX brick and leJOS NXJ supports the NXT brick. It includes a Java virtual machine, which allows Lego Mindstorms robots to be programmed in the Java programming language.

1.3 Lego MindStorms EV3

Lego Mindstorms EV3 is the third generation robotics kit in Lego's Mindstorms line. It is the successor to the second generation Lego Mindstorms NXT 2.0 kit. The "EV" designation refers to the "evolution" of the Mindstorms product line. "3" refers to the fact that it is the third generation of computer modules- first was the RCX and the second is the NXT. It was officially announced on January 4, 2013 and was released in stores on September 1, 2013.

Lego Mindstorms EV3 is a programmable set of robots produced by company Lego. Each set consists of 601 parts and, as with all Lego building blocks, the robot can be created according to our imagination [3, 6]. More parts as sensors can be purchased in addition. The basic part as well as the brain is a programmable brick and a number of sensors, which makes this robot fully programmable. For our purposes, we used a robot that can move as a tank and is equipped with a quartet of infrared sensors in the front to prevent a collision.

There are two big engines driving a crawler tread with a programmable brick attached on top. The chassis also carries a sensor head with 4 infrared sensors placed at the height of 6.5 cm from the ground (see Fig. 2). Each infrared sensor sends out an infrared signal and detects the reflections from objects ahead. The strength of the reflected signal indicates the distance from the object. The accuracy of measurement depends on object's size, colour (light colours reflect the signal better than dark), material, and other factors. The value ranging from 0 to 100 does not accurately determine the distance itself. However, values with a small deviation correspond to the values in centimetres. The sensor is incapable of detecting small distances around 1 cm from the object. The sensor in port S4 is placed on the left, S3 centre-left, S2 centre-right, S1 on the right. The left engine is connected to port B and the right circuit to port A.

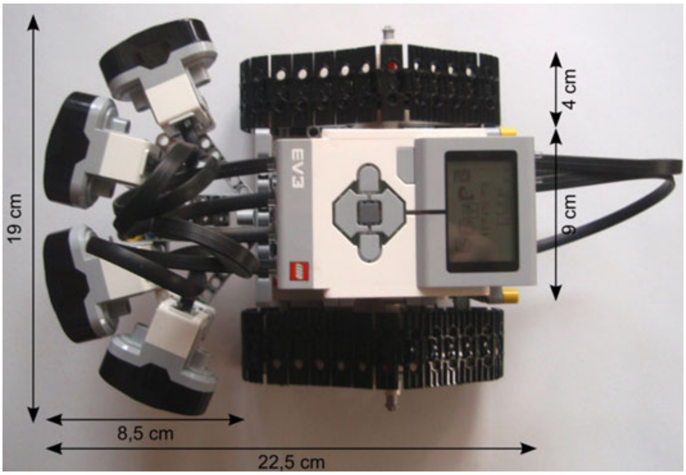


Fig. 2 The assembled robot for experimental purposes

2 Proposal and Implementation of Mobile Applications for Controlling the Robot Lego MindStorms with Verification of Adaptivity

In order that the robot could communicate through Wi-Fi, it required a Wi-Fi dongle to be connected into its USB slot. Available sources prove that a robot with such a dongle works perfectly. Before creating the project itself, it is necessary to create a PAN network which the robot connects to. The PAN network was created without a password, but when connecting the robot to the network, there were already problems to solve. The first one was setting the robot's IP address. Having entered a suitable IP address, this address did not want to display on the robot's main panel (i.e. the address was unsuccessfully created). The problem was solved with setting a default IP address, which then successfully displayed on the robot's main panel.

2.1 Creation of the Basic Layout of the Application

Class *MainActivity* is an important part of the whole application. It contains the method *onCreate*, which is automatically launched after the application start-up. It contains an entire code of how individual objects of the application should behave and display. Connection to the robot from the application was without any problems. Establishing a connection between the robot and the application works on the principle of copying the IP address and establishing the connection. After a successful connection, it was necessary to crate buttons to control the robot manually. This

Table 1 Distribution of fuzzy sets of input variables

Name	Type	LeftSupp	Kernel	RightSupp
Close	Triangle	0	0	50
Medium	Triangle	0	50	100
Far	Triangle	50	100	100

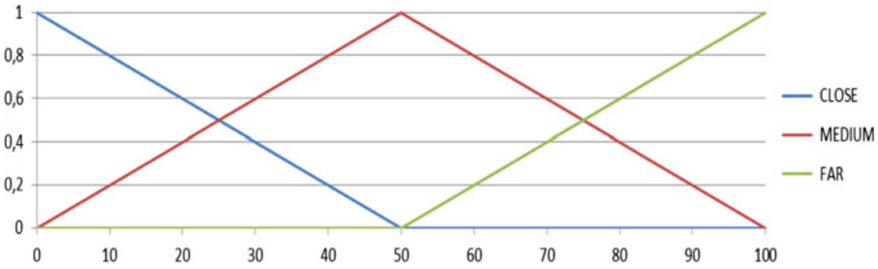


Fig. 3 Division of the fuzzy sets of input variables

phase was without any problem as well. The most important part, however, is still to be done—securing adaptive behaviour of the robot.

This library for using a fuzzy approach was created by Juan Rada-Vilela. It concerns an open-source library. **Jfuzzylite** [7] works perfectly with other classes. The base is an engine that is set according to our needs by a defuzzifier, i.e. Centre of Gravity, Mean of maximum, etc. Having created input and output variables and added them to the engine, we have to create and add fuzzy rules to the engine as well. Then we launch the engine and when we send input values to the engine, we call method *process* and get the output values. In our project, **Jfuzzylite** is to be found in the class called *FuzzyManager*.

2.2 *Input Values for Jfuzzylite*

The input values were defined as the distances that individual sensors record. There were four input variables (one for each sensor) marked as S1 to S4. The sensors can distinguish ranges between 0 and 100 cm. this range was distributed into three fuzzy sets. All input variables contain triangle-shaped fuzzy sets. These sets and their distribution are identical for each input value. Names and distribution of the sets of input variables can be seen in Table 1.

Figure 3 depicts a graphical representation of individual triangle sets.

Table 2 Distribution of fuzzy sets of output variables

Name	Type	LeftSupp	Kernel	RightSupp
Very_quick_back	Triangle	−500	−500	−375
Quick_back	Triangle	−500	−375	−250
Slow_back	Triangle	−375	−250	−125
Very_slow_back	Triangle	−250	−125	0
Stop	Triangle	−125	0	125
Very_slow_forward	Triangle	0	125	250
Slow_forward	Triangle	125	250	375
Quick_forward	Triangle	250	375	500
Very_quick_forward	Triangle	375	500	500

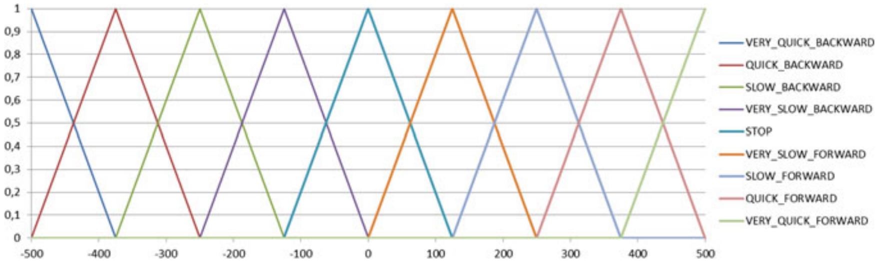
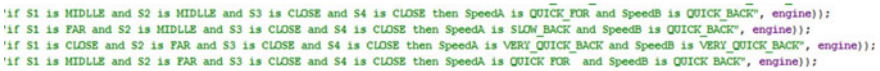


Fig. 4 Distribution of fuzzy sets of output values

2.3 Output Variables of Jfuzzylite

The output variables are represented by individual engines of the Lego MindStorms EV3 robot. One output variable for each engine. They are marked as SpeedA and SpeedB. The range was determined from 500 to 500 rpm. It was not suitable to use higher revolutions. The robot would move too fast and before the sensors recorded the distance from an obstacle and the engines reacted, the robot would hit the obstacle. Similarly to the input variable, the output variables also worked with triangle-shaped sets. Triangle-shaped sets are simple and there is higher probability that more fuzzy rules engage for given input values. The range of revolutions was equally distributed between individual fuzzy sets. Names and distribution of the sets of output variables can be seen in Table 2.

Figure 4 depicts a graphical representation of individual triangle sets of the output variables.



```
'if S1 is MIDLLE and S2 is MIDLLE and S3 is CLOSE and S4 is CLOSE then SpeedA is QUICK_FOR and SpeedB is QUICK_BACK", engine));
'if S1 is FAR and S2 is MIDLLE and S3 is CLOSE and S4 is CLOSE then SpeedA is SLOW_BACK and SpeedB is QUICK_BACK", engine));
'if S1 is CLOSE and S2 is FAR and S3 is CLOSE and S4 is CLOSE then SpeedA is VERY_QUICK_BACK and SpeedB is VERY_QUICK_BACK", engine));
'if S1 is MIDLLE and S2 is FAR and S3 is CLOSE and S4 is CLOSE then SpeedA is QUICK FOR  and SpeedB is QUICK BACK", engine));
```

Fig. 5 Sample of the rule base

2.4 Creating a Base of Fuzzy Rules

Having created the fuzzy sets for input and output variables, we can proceed to the main step, i.e. creation of the fuzzy rules base. Jfuzzylite uses classical fuzzy rules of if-then type. The antecedent contains the measured values from infrared sensors and the consequent (follows the *then* word) contains action response of individual engines. A general structure of a fuzzy rule is depicted in (1).

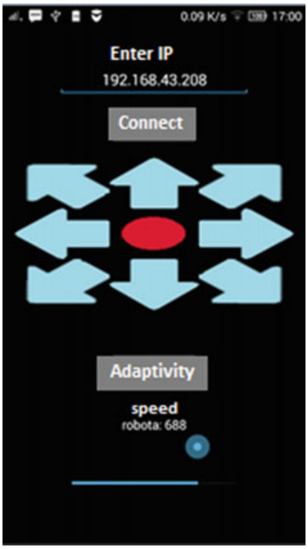
$$\boldsymbol{If}\ S1\ and\ S2\ and\ S3\ and\ S4\ is\ X\ \boldsymbol{then}\ SpeedA\ is\ Y\ and\ SpeedB\ is\ Z \tag{1}$$

In Jfuzzylite, the rules must be individually entered into the engine. As we have four input variables with three fuzzy sets, we have to assess 3^4 rules. It concerns then 81 rules containing every combination of all input variables. In order to eliminate changing individual linguistic assessments of the fuzzy sets when creating the rules, we made a simple program where we enter the linguistic assessment of the sets and the number of input variables. The program then generates all combinations of fuzzy sets for individual input variables. Thus, we had all rules and we only had to manually assess them, i.e. add the output variable. Such a created base secures that the robot behaves adaptively when going through a given environment—it does not hit obstacles that are in the given environment. A part of the rule base is presented in Fig. 5.

2.5 Instructions to Use the Proposed Application

This section is focused on a description how to use the proposed application for Lego MindStorms EV3 robot control with a verification of its adaptivity. First, create a PAN network and connect the robot into the network. When you start the application, an initial screen shows up. In its upper part, enter robot's IP address (to be found on the robot's control unit panel in its upper part) and then click the button *Connect*. If everything done according to the instructions, the robot makes a sound signal meaning that the application has been successfully connected. The button *Connect* changes its name and function to *Disconnect*.

Below the button *Disconnect*, there are eight direction buttons to control the robot. In the middle, there is a red button *Stop* to stop the robot moving. The bottom bar contains a seekbar, which enables to set the speed of the robot (max. 900 rpm), but only in manual mode. Above the seekbar, there is button *Adaptivity* to launch the adaptive mode. If a robot is idle and we press this button, the robot starts behaving



Enter IP
192.168.43.208

speed
robota: 688

Fig. 6 Initial screen of the application

adaptively in the environment it is in. Repeated pressing the button means switching from an adaptive mode to a manual mode. The initial screen of the application is in Fig. 6.

3 Experimental Part and Verification of Adaptivity

The objective of this section is to verify adaptive behaviour of the robot in a given environment using a fuzzy rules base. The proposed application has been tested on several devices. It concerned, among others, Lenovo P70 with display resolution 1280 × 720, tablet Samsung Galaxy Table 2 with display resolution 1024 × 600. Both devices ran the application with no problems and the adaptive behaviour of the robot worked as well.

3.1 Experiments and Definition of Criteria

The robot was placed into a square of 15 × 15 centimetres. Its objective was to go through the created environment without hitting an obstacle or getting stuck on a place. For each experimental environment, the following data was gathered: number of collisions (direct hit into an obstacle), number of getting stuck on a place, time to get from A (marked with a red circle) to B (marked with a green circle), and robot's

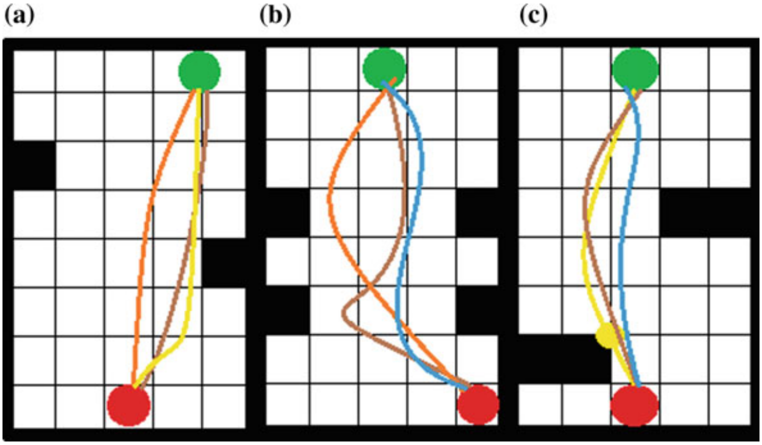


Fig. 7 Selected experiment and individual environments, from left A, B and C

Table 3 Results of the experiment

	Collisions	Deadlocks	Time (s)
Transit 1/A	0	0	9
Transit 2/A	0	0	10
Transit 3/A	0	0	11
Transit 1/B	0	0	10
Transit 2/B	0	0	14
Transit 3/B	0	0	10
Transit 1/C	0	0	11
Transit 2/C	0	0	10
Transit 3/C	0	0	9

trajectory. Robot's speed was set to −500 to 500 rpm. Scanning of the environment is done every 0.5 s.

The first three experimental environments had dimensions 120 × 75 cm. Each had a different layout of obstacles to be avoided. The robot had to go through the route from A to B with no hits. There were three passes through each environment with recording of robot's route. Moreover, in each variant the robot was turned to a different direction then only directly to the destination B. In the first variant (A), the sensors headed to the left (the same as in environment B). In environment C, the sensors were turned to the right, thus the robot had to turn. Individual experimental environments can be seen in Fig. 7.

When going through the C variant, the turning robot touched the edge of an obstacle with its cable, which is not considered to be a direct hit (a yellow mark) and thus is not counted in the table below. The measured values can be found in Table 3.

In all passes, the robot got from A to B and it did not directly hit any obstacle or got stuck. In general, we consider the created rule base and the final application as satisfactory because the robot did not hit obstacles in the adaptive mode. The manual mode of the robot also corresponds with the original intentions of robot control.

4 Conclusion

The developed application and its parts run with no problems on more devices supporting Android operating system. The application perfectly functioned on the created experimental environments, including the adaptive mode. This experiment has brought a lot of experience concerning programming in operating system LeJOS and its integration with robot Lego MindStorms EV3 and Android. The proposed application is already used in lessons at University of Ostrava as a supportive tool to create exercises in course Application of Artificial Intelligence. Students continue to develop this application and in future we expect more applications to be created for mobile devices and their implementation into lessons.

Acknowledgements The research described here has been financially supported by the University of Ostrava grant SGS04/PrF/2018.

References

1. Åström KJ, Wittenmark B (2013) Adaptive control. Dover Publications, Mineola, New York
2. Tzafestas SG (2013) Introduction to mobile robot control. Elsevier, London
3. Bell M, Floyd J, Kelly JF (2017) LEGO mindstorms EV3. Apress, Berkeley, CA
4. Grandi R, Falconi R, Melchiorri C (2014) Robotic competitions: teaching robotics and real-time programming with LEGO mindstorms. In: Proceedings of 19th world congress, the international federation of automatic control Cape Town, South Africa (2014)
5. Burnette E (2009) Hello, Android: introducing Google's mobile development platform (Pragmatic Programmers). The Pragmatic Bookshelf, Raleigh, North Carolina
6. Konvicka J (2014) Lego robot control using mobile devices, Bachelor thesis, University of Ostrava, Ostrava
7. JFuzzyLite (2015) http://www.fuzzylite.com/java/. Accessed 20 Oct 2015

Improving Medical Short Text Classification with Semantic Expansion Using Word-Cluster Embedding

Ying Shen, Qiang Zhang, Jin Zhang, Jiyue Huang, Yuming Lu and Kai Lei

Abstract Automatic text classification (TC) research can be used for real-world problems such as the classification of in-patient discharge summaries and medical text reports, which is beneficial to make medical documents more understandable to doctors. However, in electronic medical records (EMR), the texts containing sentences are shorter than that in general domain, which leads to the lack of semantic features and the ambiguity of semantic. To tackle this challenge, we propose to add word-cluster embedding to deep neural network for improving short text classification. Concretely, we first use hierarchical agglomerative clustering to cluster the word vectors in the semantic space. Then we calculate the cluster center vector which represents the implicit topic information of words in the cluster. Finally, we expand word vector with cluster center vector, and implement classifiers using CNN and LSTM respectively. To evaluate the performance of our proposed method, we conduct experiments on public data sets TREC and the medical short sentences data sets which is constructed and released by us. The experimental results demonstrate

Y. Shen · Q. Zhang · J. Zhang · K. Lei
Institute of Big Data Technologies Shenzhen Key Lab for Cloud Computing Technology & Applications School of Electronics and Computer Engineering (SECE), Peking University, 518055 Shenzhen, People's Republic of China
e-mail: shenying@pkusz.edu.cn

Q. Zhang
e-mail: zhangqiang@sz.pku.edu.cn

J. Zhang
e-mail: 1701213660@sz.pku.edu.cn

K. Lei
e-mail: leik@pkusz.edu.cn

J. Huang · Y. Lu (✉)
ShenZhen Key Lab for Visual Media Processing and Streaming Media, Shenzhen Institute of Information Technology, 518172 Shenzhen, People's Republic of China
e-mail: luyuming@sziit.edu.cn

J. Huang
e-mail: 1701213596@sz.pku.edu.cn

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_41

that our proposed method outperforms state-of-the-art baselines in short sentence classification on both medical domain and general domain.

Keywords Text classification · Word-cluster embedding · Hierarchical agglomerative clustering

1 Introduction

Short text classification has been proven promising in natural language processing tasks, such as social network sentiment analysis [1], product review classification [2] etc. However, the medical short texts usually contain synonym, alias and acronym but lack contextual information, leading to semantic ambiguity.

Moreover, there are many oral expressions or imprecise descriptions in the medical field. For example, acquired immune deficiency syndrome is usually abbreviated as AIDS. According to the CBOW model, the vector of the target word is predicted based on context vectors. These words have semantic similarity and often appear in the same context. Therefore, they have similar word vectors and are close in semantic space. In the semantic space, similar words are more likely to be grouped together and different words are more likely to be aggregated into different clusters.

To alleviate these problems, we propose the cluster-based word vector expansion method, which uses Hierarchical Agglomerative Clustering (HAC) to cluster the words by calculating the distance between words in the semantic space. Concretely, it calculates the cluster center vector as the potential theme information of words in this cluster. Then it adds cluster center vector to the corresponding word vector to expand the semantic features of words in short text. Clustering algorithm is an unsupervised machine learning algorithm, which divides samples into different groups based on the features. The scope of the group is not clear, because there not exist ground truth class assignments. Clustering aims to make the samples in the same group as similar as possible and the samples in different groups as different as possible.

With the cluster-based word vectors, the short text classification is carried out through Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) neural networks respectively. Our method is evaluated on two datasets: The Chinese Medical Short Sentence (CMSS) corpus we developed and released and TREC corpus in common field. The experiments demonstrate that our method has robust superiority over competitors and sets state-of-the-art. The main contributions of this article include:

1. We propose cluster-based semantic expansion method to reduce the problems of low classification accuracy caused by feature sparseness and semantic ambiguity in medical short text classification field.
2. Based on the hierarchical aggregation clustering algorithm, we calculate the cluster embedding which represents the implicit topic information of the cluster.

3. The Chinese Medical Short Sentence (CMSS) corpus we develop and release contains 17,787 sentences that classified in three symptom severity rating, which is slightly, moderately and heavily.

The rest of the article is organized as follows. Section 2 introduces the related work about word representation and short text classification. Section 3 gives a detailed description of the overall framework of our method. Section 4 presents experimental setup, results and analysis, including the construction of Chinese medical dataset, the experiment in medical dataset and experiment in published TREC dataset. Section 5 summarizes this work and the future direction.

2 Related Work

The accuracy of short text classification is affected by sparseness of features since short text usually contains fewer characters. Bag-of-words (BoW) model can not be directly applied to short text representation since it expresses words using high-dimensional and sparse vectors, and ignores the order and semantic relations between words [3]. According to the distributed hypothesis, words appearing in the same context often have similar semantics. Based on the distributed hypothesis and the neural network model, Mikolov et al. [4] proposed Word2Vec model which including the Skip-Gram model and CBOW (Continuous Bag-of-Words) model. Each of the Skip-Gram and CBOW method defines a method for creating an unsupervised learning task from plain raw corpora. The CBOW model trains each word against its context, while Skip-Gram trains each the context against the word. Pennington et al. [5] proposed the global vectors for word representations (GloVe). The global log-bilinear regression model combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. The Word2Vec and GloVe models are widely used, since they consider the contextual information to obtain dense low-dimensional real-valued vectors and thereby overcome the shortcomings of BoW model.

To improve the accuracy of short text classification, many studies adopted deep neural network for the short texts classification. Facebook researchers Joulin et al. [6] proposed FastText model to represents sentence vectors using the mean of word vectors. Kim et al. [7] proposed the two-channel CNN using both statically and dynamically updated word embedding as input. As we all know, CNN could only capture local features, but ignore long-distance dependencies between words. Socher et al. proposed Recursive Neural Network (RNN) model for the sentiment classification [8]. It cannot obtain satisfied classification accuracy due to the vanishing gradient problem. Hochreiter S. et al. [9] employed LSTM to solve the aforementioned problem by replacing a single unit with a more complex memory unit. Zhou et al. [10] proposed C-LSTM model to represent and classify sentences. This model extracts high-dimensional phrase vectors using convolutional layer, then feeds phrases vec-

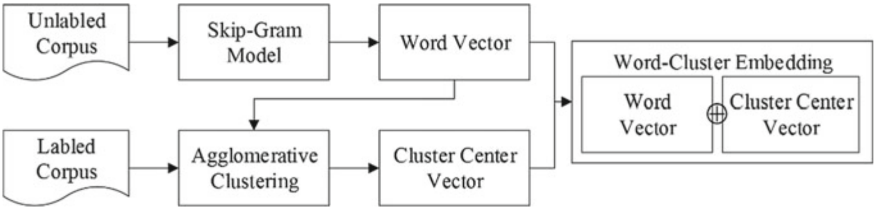


Fig. 1 Framework of cluster-based semantic expansion model

tors into LSTM layer to get sentence vector. The model performs well since captures both local features of the phrase and the global features of sentence.

In the use of clustering, Song et al. [11] proposed the cluster-based multiple SVM classifiers to classify polarity of short product reviews. Wang et al. [12] proposed the method to classify short texts based on semantic clustering and CNN model. This method finds semantic clusters based on searching density peak, and uses n-gram to detect candidate semantic units in short text.

Combine the advantages of the aforementioned methods, we propose the cluster-based semantic expansion method to improve the performance of short text classification in medical field. We perform experiment on both CMSS and TREC data sets.

3 Methodology

3.1 Cluster-Based Semantic Expansion Model

We propose a cluster-based semantic expansion model to reduce the problem of sparse features in short text. As shown in Fig. 1, we first obtain word embedding from unlabeled medical corpus through unsupervised Skip-Gram model. Then, we use hierarchical agglomerative clustering to cluster words in labeled corpus. Finally, we obtain word-cluster embedding through the concatenation of word embedding and cluster embedding.

Word Embedding. Skip-Gram is used to train the context of words to obtain word embedding based on the unlabeled medical corpus. The Skip-Gram's objective function which should be maximized is presented in Eq. (1):

$$\mathrm{J}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

In Eq. (1), w_t and w_{t+j} indicate to the middle and context words separately, and c points to the size of training window. Given the word w_t in the middle, we can compute the log probability of predicting word w_{t+j} from $-c$ to c in the training window. The

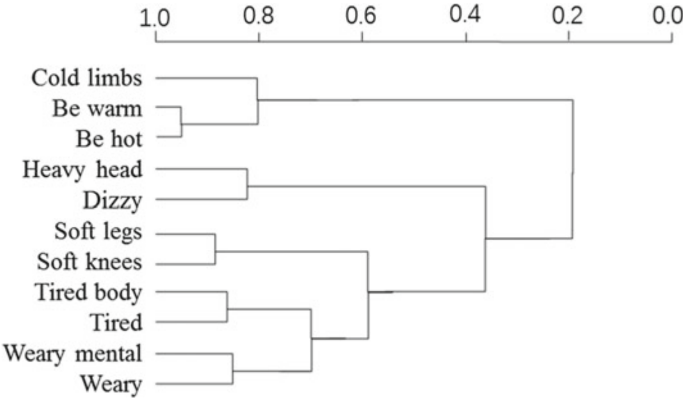


Fig. 2 The example of HAC in medical field

value of $p(w_{t+j}|w_t)$ can be calculated by Eq. (2), where v_{w_t} and $v_{w_{t+j}}$ represents the distribution representations of the middle and context words respectively. Through Eqs. (1) and (2), we can finally obtain the low-dimensional dense word embedding.

$$p\left(w_{t+j}|w_t\right) = \frac{\exp\left(v_{w_{t+j}}^T v_{w_t}\right)}{\sum_{w=1}^{W} \exp\left(v_w^T v_{w_t}\right)} \tag{2}$$

Hierarchical Agglomerative Clustering. Hierarchical Agglomerative Clustering (HAC) starts with every single sample in a single cluster, then iteratively merges two most similar clusters until there is only one cluster or preset number of clusters. HAC tends to produce smaller clusters with reasonable preset number of clusters. Therefore, it is suitable for word clustering which includes fewer words per cluster. As shown in Fig. 2, HAC produces tree structure in bottom-up direction.

The selection of similarity function is critical. The common similarity functions include single linkage, complete linkage and average linkage. We adopt the average linkage by taking into account the sensitivity to outlier, global quality and time complexity. The other two functions are deprecated due to two reasons. (1) The single linkage calculates similarity by the distance between nearest samples in two clusters, so the cluster has good local consistency but with poor global quality. (2) The complete linkage produces compact clusters by calculating the similarity based on the distance between furthest samples in two clusters. Complete linkage is sensitive to outliers.

Given two clusters A and B, their similarity can be calculated by Eq. (3):

$$sim(A, B) = \frac{\sum_{u \in A, v \in B} sim(u, v)}{size(A) * size(B)} \tag{3}$$

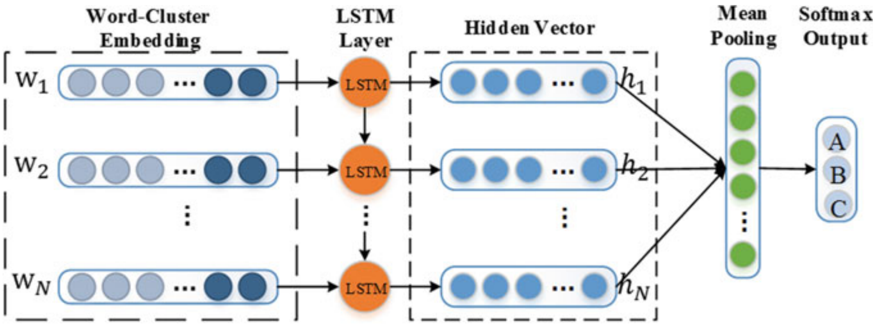


Fig. 3 The classifier using LSTM model

Here, sim(u, v) represents the similarity between the samples u and v. As shown in Eq. (4), the sim(u, v) represents the Euclidian distance-based similarity:

$$sim(u, v) = \frac{1}{1 + \sqrt{\sum_{j=1}^{n} \left(u_j - v_j\right)^2}} \tag{4}$$

We calculate the cluster embedding of each cluster by Eq. (5). Here, m represents number of words in a cluster, while V_i represents the vector of the ith word.

$$\mathrm{C} = \frac{1}{m} \sum_{i=1}^{m} V_i \tag{5}$$

We alleviate the problem of ambiguous synonyms and feature sparseness in short text by the application of cluster embedding. The cluster embedding represent the implicit topic of all words in a cluster.

3.2 Short Text Classifiers

We use CNN and LSTM models to classify the short texts. Take LSTM as an example (see Fig. 3), the first layer of the network is word embedding layer which transforms words into representations that capture syntactic and semantic information about the words. Each word is converted into a real-valued vector. Therefore, the input to the next layer is a sequence of real-valued vectors. We choose Word2Vec that is pre-trained on a corpus containing medical texts. Average-pooling is adopted to conduct pooling operations. The Softmax layer outputs the probabilities of short texts belonging to different categories.

Table 1 The medical corpus used to train skip-gram model

Corpus	Words(million)
Baidu encyclopedia (medical)	37.8
Hudong Encyclopedia (medical)	39.2
Chinese Wikipe	44
General EMR	312.7
Stroke EMR	1.2
7th edition of internal medicine	0.4

[a] https://baike.baidu.com/science/medical
[b] http://fenlei.baike.com/%E5%8C%BB%E5%AD%A6/
[c] http://download.wikipedia.com/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2

The CMSS data sets we developed is classified in three symptom severity rating which is slightly, moderately and heavily. The greater the probability is, the more likely the symptom severity rating it belongs to.

4 Experiment

4.1 Datasets

Word Embedding of Medical Terms. Five medical data sets and one open-domain data set is adopted in our study. As shown in Table 1, the general EMR (112,262 medical records) contains 312.7 million words, the stroke EMR (4,588 medical records) contains 1.2 million words, and the medical textbook (the 7th edition of Internal Medicine) contains 400,000 words. We obtain 490,000 word vectors using Skip-Gram model based on the aforementioned data sets.

To solve the Out-Of-Vocabulary problem, we simply use the zero-valued vector to represent the words not appearing in the training corpus. The long term in medicine can lead to the emergence of word segmentation problems. For example, "器质性脑病综合症" could be mistakenly divided into three words, "器质性", "脑病" and "综合症". Segmentation problems may cause ambiguity and decrease the accuracy of short text classification. To this end, we construct a medical dictionary which contains 21,483 words based on ICD-10 Disease Codes and Sogou Medical Thesaurus. Jieba segmentation tool is employed for the medical corpus word segmentation with the help of the medical dictionary we construct.

Chinese Medical Short Sentence. In this paper, we focus on short text classification in the medical field in Chinese language. We obtain 2,415 short sentences from the *Guiding Principles of New Clinical Drug*, then we extend the corpus with the Chinese Wikipedia redirect dictionary that contains 640,000 synonym pairs. If a word in short sentence exists in the Wikipedia redirected vocabulary, the original word will be replaced by its synonym, and therefore a new short sentence is generated. For example, the short sentence “slightly flu” could be converted into “slightly cold” since “flu” and “cold” are synonyms.

Based on these short sentences, we construct and release a Chinese Medical Short Sentence (CMSS) corpus, to make our work more reproducible. This corpus contains a total of 17,787 sentences, among which, the sentences related to symptom severity rating slightly, moderately and heavily are 5,263, 6,072 and 6,452 respectively.

4.2 Implementation Details

We use word-cluster embedding as input matrix, and use CNN and LSTM model to classify short text. The CNN model is operated with two convolutional layers, two max pooling layers and one fully connected layer. Considering the training efficiency, the number of convolution kernels is set to be 64, the kernel size is set to be 5, the pooling window size is set to be 2, and the batch size is set to be 128. The LSTM is operated with one LSTM layer, one mean-pooling layer and one fully connected layer. The number of LSTM cell is set to be 300, while the batch size is set to be 128.

4.3 Experiment Based on Medical Dataset

Experiments Based on Different Cluster Algorithms. We employ HAC to cluster 2,718 words in CMSS data sets, and use Grid Search to find the decent number of clusters in range of 200–2000. To compare the performance of different clustering algorithms in short text classification, we adopt word embedding without cluster-based expansion as input matrix in different baseline models. We perform three different experiments separately using Affinity Propagation Clustering (APC), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and HAC to expand word embedding. Finally, we use CNN and LSTM model to extract features and classify short text.

As shown in Fig. 4, both CNN and LSTM model using HAC-based word-cluster embedding achieve the best results. The reason may lie in: (i) the HAC clusters words in bottom-up direction, which can produce small clusters when given reasonable number of clusters. Therefore, it is suitable for word clustering which usually contains fewer words in a cluster. (ii) we use the average linkage to calculate the similarity between clusters, which is not sensitive to outliers.

Experiments Based on Different Classification Algorithms. To evaluate the performance of our cluster-based word-cluster embedding, we compare our method with baselines in CMSS data sets. LibLinear [13] is an open source library for large-scale linear classification, which supports logistic regression and support vector machines. RCNN [14] uses recurrent structure to capture contextual information and max pooling layer to capture key components in text. C-LSTM [10] utilizes CNN to

Fig. 4 Experimental results based on different cluster algorithms

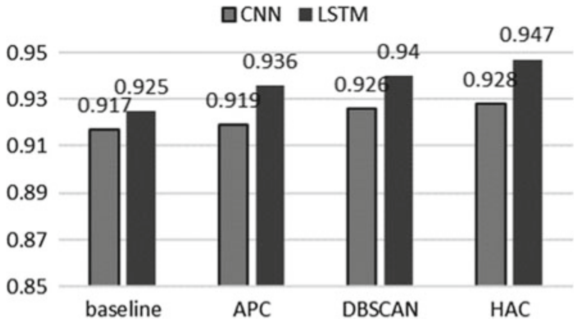


Table 2 Experimental results based on different classification algorithms

Model	Accuracy	Reported in
LibLinear	0.908	Fan et al. 2008
C-LSTM	0.929	Zhou et al. 2015
RCNN	0.933	Lai et al. 2015
HAC-CNN	0.928	Our method
HAC-LSTM	**0.947**	**Our method**

extract higher-level phrase representations, and employs LSTM to obtain the sentence representation.

As shown in Table 2, the HAC-CNN model achieve similar accuracy to C-LSTM, but slightly poor compared with RCNN. The HAC-LSTM model outperforms all baselines.

4.4 *Experiment Based on Common Dataset*

To further evaluate the performance of our method, we perform experiments on public data sets: TREC. The TREC released by the UIUC Cognitive Computation Group contains 6 categories with over 6,000 labeled data. To obtain word embedding vectors, we use Google's open-source word vector trained in Google News (about 100 billion words), which contains roughly 3 million words and phrases, each of 300 dimensions.

We use HAC to cluster words and adopt cross-validation to find decent number of clusters. The experimental results show that model perform best when the number of clusters is 1000. We employ the cluster-based word-cluster embedding as input matrix, and use CNN and LSTM model to extract features and classify sentences.

We use the following baselines: CNNs-non-static [7] uses word vectors fine-tuned during the training, and CNNs-multichannel both uses static and non-static word vectors. TFIDF+SVMs model [12] classifies the sentences using SVM classifiers based on the Term Frequency (TF) and Inverse Document Frequency (IDF) of each word in a sentence. Semantic-CNN [12] expands word embedding based on density

Table 3 The results of short text classification on TREC data sets

Model	Accuracy	Reported in
SVMs	0.95	Silva et al. [15]
TFIDF + SVMs	0.943	Wang et al. [12]
CNNs-non-static	0.936	Kim [7]
CNNs-multichannel	0.922	Kim [7]
DCNN	0.93	Kalchbrenner et al. [16]
DCNNs	0.956	Ma et al. [17]
Tree CNN	0.96	Komninos and Manandhar [18]
Semantic-CNN	0.956	Wang et al. [12]
C-LSTM	0.946	Zhou et al. [10]
RCNN	0.96	Lai et al. [14]
HAC-LSTM	**0.977**	**Our method**

peak clustering, and classify sentences using multi-scale CNN. SVMs [15] model classifies short text using SVM classifiers based on 60 features artificially extracted such as unigrams, bigrams, POS tags, syntax analysis, ephemera, WordNet and so on. DCNN uses the dynamic k-max pooling layer [16] to extract global features. DCNNs uses n-gram information from dependent trees [17] to capture long-range dependencies. Tree CNN [18] uses the information from the dependent tree to enrich the semantic information of word embedding.

As shown in Table 3, our model outperforms the state-of-the-art on the TREC data sets. The improvement is mainly due to the cluster-based semantic expansion method, which reduces the problems of low classification accuracy caused by feature sparseness and semantic ambiguity.

5 Conclusion and Future Work

In this study, we propose a cluster-based semantic expansion method based on hierarchical agglomerative clustering, which effectively incorporate word embedding into cluster embedding to obtain more semantic information. Experimental results on two benchmark datasets demonstrate the superiority of our proposed method on short text classification task. We believe that the most promising avenues for future research include experimenting with methods of named entity recognition method to improve the word segmentation.

References

1. Dos Santos CN, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts COLING, pp 69–78
2. Buschmeier K, Cimiano P, Klinger R (2014) An impact analysis of features in a classification approach to irony detection in product reviews. WASSA@ ACL
3. Sriram B, Fuhry D, Demir E et al (2010) Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, pp 841–842
4. Mikolov T, Chen K, Corrado G et al (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
5. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
6. Joulin A, Grave E, Bojanowski P et al (2016) Bag of tricks for efficient text classification. arXiv:1607.01759
7. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv:1408.5882
8. Socher R, Perelygin A, Wu J et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1631–1642
9. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
10. Zhou C, Sun C, Liu Z, Lau F A C-LSTM neural network for text classification. arXiv:1511.08630
11. Song J, He Y, Fu G (2015) Polarity classification of short product reviews via multiple cluster-based SVM classifiers. PACLIC
12. Wang P, Xu J, Xu B et al (2015) Semantic clustering and convolutional neural network for short text categorization. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), vol 2, pp 352–357
13. Fan RE, Chang KW, Hsieh CJ et al (2008) LIBLINEAR: a library for large linear classification[J]. J Mach Learn Res 9:1871–1874
14. Lai S, Xu L, Liu K et al (2015) Recurrent convolutional neural networks for text classification. In: AAAI, vol 333, pp 2267–2273
15. Silva J, Coheur L, Mendes AC et al (2011) From symbolic to sub-symbolic information in question classification. Artif Intell Rev 35(2):137–154
16. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. arXiv:1404.2188
17. Ma M, Huang L, Xiang B et al (2015) Dependency-based convolutional neural networks for sentence embedding. arXiv:1507.01839
18. Komninos A, Manandhar S (2016) Dependency based embeddings for sentence classification tasks. HLT-NAACL, pp 1490–1500

A Deep Learning Approach to Mining the Relationship of Depression Symptoms and Treatments for Prediction and Recommendation

Juntian Lin, Guan Luo, Zhu Zhan and Xiaoyao Guan

Abstract Background: Behavior regulation and clinical intervention have a significant effect on depression treatments. This study aims to make a comparison between behavior regulation and clinical intervention for depression based on a large-scale dataset. Methods: We collect user-reported data from an online survey tool including depression symptoms, treatments and effectiveness of treatments (n = 91873). A deep learning approach is used to build an effective model to evaluate the effects on treatment methods for depression. The Skip-gram model is chosen to generate meaningful vector representations of symptoms and methods. Precision, recall and F1 score are calculated to evaluate the model performance. Results: Unidirectional model achieves higher F1 score than non-unidirectional model (0.71 vs. 0.63). The behavior regulation is better than the clinical intervention for mild depression symptoms. However, the clinical intervention for moderate or severe depression symptoms has obvious advantages. Conclusions: These experiments prove that the symptoms have unidirectional influence on the choice of regulatory methods. The behavior regulation and clinical treatment have different advantages for depression. These findings could help clinicians to choose better depression treatments.

Keywords Depression · Deep learning · Behavior regulation

J. Lin · G. Luo (✉) · Z. Zhan · X. Guan
National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
e-mail: gluo@nlpr.ia.ac.cn

J. Lin
e-mail: juntian.lin@ia.ac.cn

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_42

1 Introduction

Depression (Major depressive disorder) is a common but serious mental disorder. The World Health Organization (WHO) shows that more than 300 million people in the world are affected by depression [1]. Persistent feelings of sadness and lack of interest or pleasure in activities are defined to be the most important features of depression in the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V). Besides the two above symptoms, DSM-V lists the other seven typical depression symptoms (Table 1).

The depression treatment can be divided into antidepressant and psychology. Antidepressant is the most common choice for depression treatment. However, it cannot treat depression effectively because of its poor compliance and persistence. Sansone and Sansone [2] suggest that approximately half of the depressive patients cannot persist in antidepressant treatment. To enhance the efficacy of depression treatment, psychological treatment is usually used as an important complement to antidepressant. Psychology can also improve the quality of life assessment [3].

In addition to the above clinical treatments, the use of behavior regulation, which contains energy regulation, emotion regulation and so on, is also regarded as an important aspect of recovery for depression [4]. Patients with higher level of depressive symptoms prefer using rumination and suppression, rather than reappraisal [5]. Many studies focus on the mechanism by which behavior regulation affects depression. However, the public data on the efficacy of self-regulating is small.

The aim of this study is to examine the effect of behavior regulation and clinical intervention for depression on a large-scale dataset. We build a model to predict whether the symptoms of the participants can be alleviated, so that we can use this model to evaluate the effect of each method in any condition.

Table 1 Typical depression symptoms (DSM-V)

Symptom	Description
A	Depressed mood or irritable nearly every day
B	Diminished interest or pleasure in activities
C	Significant weight loss (5%) or change in appetite
D	Insomnia or hypersomnia
E	Psychomotor agitation or retardation
F	Fatigue or lack of energy
G	Feelings of inappropriate guilt or worthlessness
H	Decreased ability to concentrate, or indecisiveness
I	Suicidality

2 Related Work

Many studies have shown that behavior regulation and clinical intervention have a significant effect for depression treatment [6–8]. These studies form the basis of making comparisons on different methods through a large-scale data. Martin and Dahlen [9] show that emotion regulation strategies, especially positive reappraisal and rumination, is valuable factors to predict negative emotions like depression. Nevertheless, depressed people cannot decrease negative feelings in consequence of using emotion regulation strategies [10]. These studies imply that methods are affected by depression symptoms. We use a deep learning approach to enrich the information about methods and to compare behavior regulation with clinical intervention.

3 Method

3.1 Dataset Description

We use an online survey tool to collect data from voluntary participants. The dataset eventually contains 91873 valid data. Three aspects of data are recorded for every participant, i.e. depressive symptoms, methods and effect. The Patient Health Questionnaire (PHQ-9) is a self-report inventory designed for depression screening [11]. We use PHQ-9 to obtain nine depressive symptoms, denoted as letter A-I. Each depression symptom is divided into four levels: “Not at all”, “Several Days”, “More than half the days”, “Nearly every day”, and is recorded as 0–3, like A1 means “Depressed mood in several days”. At the same time, we list eleven commonly used depression adjustment methods, denoted as a-k, for participants to follow, and evaluate the effect of the chosen methods (0-good, 1-bad) (Table 2).

Table 2 Depression adjustment methods

Method	Description
a	Hot showers
b	Exercising
c	Talking
d	Suppression
e	Journey
f	Eating
g	Reading
h	Sex life
i	Shopping
j	Reappraisal
k	Seeking medical advice

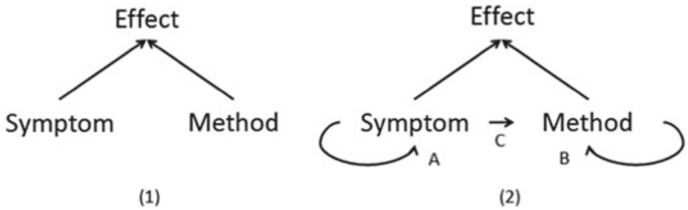


Fig. 1 Research approaches to depression

All eligible participant should answer all of questions independently. Considering that the purpose of this study is to predict the result of a combination of methods for a patient with some depression symptoms, we drop out those data with no symptoms or no methods. At last, we have 91873 valid records. Participants are cognitively healthy men and women aged in 18–70. The depression mean score for males (n = 27018) is 9.90 ± 6.54 and the mean score for female (n = 64855) is 9.92 ± 6.36. There is no significant difference in PHQ-9 score between males and females ($t = -0.58$, $df = 49281$, $p = 0.57$). More than half of participants (52.11%) feel that those methods can decrease their depressive symptoms.

3.2 *Study Design*

To predict the patient's condition is improved or not after he/she uses the clinical or self-regulating methods, a common approach in Fig. 11 is to treat each depression symptom and method as an independent feature and then use a classification algorithm to solve the problem. This structure is simple but it does not consider the interaction between features. Thus, we design a more reasonable structure to describe the relationship among symptoms, methods and effect as shown in Fig. 12.

First, different symptoms and methods are correlated in their respective sets. Second, the symptoms have influence on the choice of regulatory methods [12]. The first problem can be solved by using the traditional interactive terms in Fig. 11. However, the second one cannot do this because the interaction terms cannot describe the unidirectional relationship. Since depression symptoms have effects on the chosen of methods, the opposite does not hold true. In order to model the unidirectional relationship between symptoms and methods, we propose a word embedding approach to model the relationship.

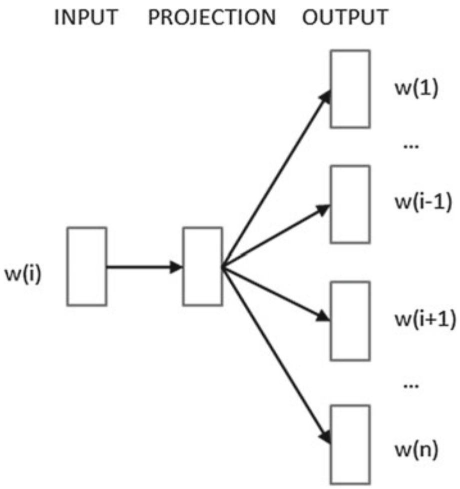


Fig. 2 Skip-gram model structure

3.3 Learning Symptom and Method Vectors

We treat the symptoms and methods in PHQ-9 as words, and the combination of individual symptoms and methods as a sentence. The efficacy of the methods is the learning target. Thus, by training a neural network using the word embedding algorithm (Word2Vec), we learn the interrelationships between symptoms and methods. In particular, we make an improvement on the Word2Vec's algorithmic to learn unidirectional information between symptoms and methods (Fig. 2).

Mikolov [13] shown that Skip-gram model in Word2Vec can capture more semantic information than Continuous Bag-of-Words(CBOW). Hence, in our approach we use Skip-gram model to learn symptom and method vectors. In symptom vector training process, we select one sample at each epoch, then we take turns choosing depression symptom w(i) as input and the rest of the symptoms as the model output w(1),…, w(i − 1), w(i + 1),…, w(n), where n is the number of symptoms in this sample. The method vector training process is the same.

The difference between our approach and the Skip-gram prototype lies in the fact that our vector embedding models do not have the concept of a sequence. To reduce the sensitivity of the model to these orderings, we shuffle the symptoms and methods each time in training process. Since there are not many symptoms or methods to choose from, the native loss is calculated directly.

3.4 Model Building and Evaluation

For classification, we use the same linear softmax classifiers as fastText to classify the samples [14]. For a training sample, we summed up the symptom vectors as features

Table 3 Algorithm performance for different learning methods

	Precision	Recall	F1 score
Logistic regression	0.61	0.60	0.60
Non-unidirectional	0.69	0.58	0.63
Unidirectional	0.75	0.67	0.71

of the symptom part. Similarly, the method vectors were summed up as the features of the method part, and then the two vectors were combined as the final sample features. The dataset is randomly divided into training set and test set according to 7:3.

We use precision, recall, and F1 scores to evaluate the performance. In a classification task, the test set could be divided into four types as follows: true positives (TP), false positives (FP), true negatives (TN), false negatives (FN). If the effect of treatment is proven present in a participant, the given model also indicates the effect of treatment in this participant, the result of the model is considered true positive. The other three types are similar to the definition of TP. Precision is the fraction of TP in the sum of TP and FP. Recall is the fraction of TP in the sum of TP and FN. F1 score is the harmonic mean of precision and recall.

4 Result

4.1 Learning Unidirectional Relationship

We choose logistic regression method as the baseline model to compare the performance. To learn the symptom and method vectors, the dimensionality of symptom and method vectors is set to 20. The results are shown in Table 3.

We see that the F1 score is 0.60 for baseline model, and 0.63 for non-unidirectional model. These two models are closer in the F1 score. Method vectors added symptoms information during the training and the F1 score is 0.71 for unidirectional model. The result indicates that the method is affected by the symptoms.

4.2 Symptom Vector Analysis

By computing the cosine similarity between different symptoms, the symptoms of depression show a stratification characteristic, mild ("Several Days"), moderate ("More than half the days") and severe ("Nearly every day") symptoms are clustered based on the severity of symptoms. For example, Table 4 shows A1/A2/A3 and its nearest nine symptoms.

Table 4 The nearest symptoms (context) to symptom A (center)

Center	Context
A1	B1, F1, G1, H1, C1, E1, D1, I1
A2	C2, G2, F2, B2, D2, E2, H2, I2
A3	D3, F3, E3, C3, G3, B3, H3, I3

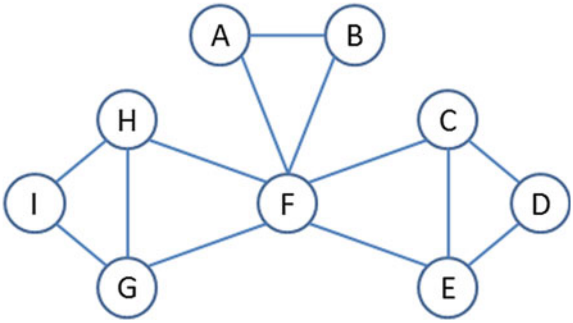


Fig. 3 The mild symptom network

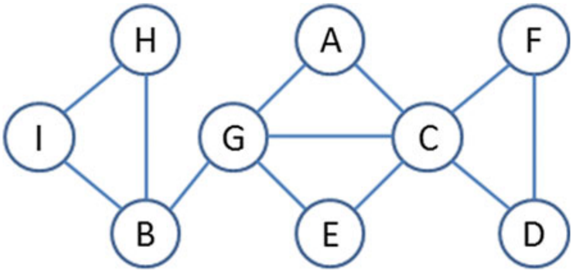


Fig. 4 The moderate symptom network

Based on the fact of stratification, we construct the symptom networks for the symptoms of grade1, grade2, and grade3 respectively. The symptom network has nine vertices corresponding to nine symptoms, the edge between vertices means the symptoms relationship. The symptom network is an undirected graph because there is no direction for similarity.

The symptom network is established as follows: the network vertices are nine depressive symptoms and for every symptom 'X', the other two symptoms 'Y' and 'Z' which have the two highest cosine similarity scores symptoms are selected to create edges. Finally, the symptom networks are shown in Figs. 3, 4 and 5, mild (Fig. 3), moderate (Fig. 4) and severe (Fig. 5).

The mild symptom network consists of one core symptom 'F' and three groups 'AB', 'CDE', 'GHI'. The moderate symptom network has no core symptom, one core group 'ACEG' and two subgroups 'DF' and 'BHI' are made up of this network. The severe symptom network consists of two completely independent networks.

First of all, we calculate the average degree of three symptom networks. They are 2.78, 2.67, 2.22, respectively from mild to severe of the symptom score. The link

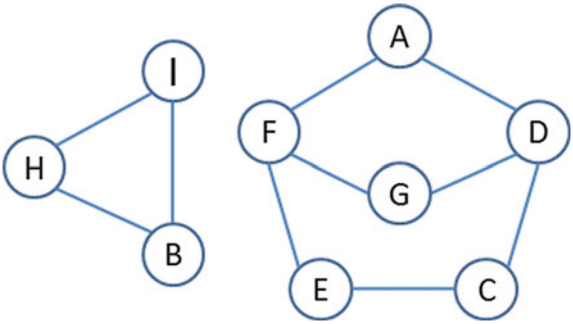


Fig. 5 The severe symptom network

density of the symptom network is reducing. This phenomenon indicates that the relationship between the symptoms become more concentrated with the increasing of symptom scores. At the same time, the maximum degree of the three symptom networks are 6, 5, 3, which is reducing with the increasing of severity of symptoms. This phenomenon reflects the disappearance of the central symptoms in depression network. In other word, the more severe the symptoms, the greater discrepancy between participants.

Second, the development of symptoms ‘B’, ‘G’, ‘H’, and ‘I’ is interesting. In the mild symptoms network, symptom ‘B’ is mainly related to symptoms ‘A’ and ‘F’. The symptoms ‘A’ and ‘B’ are the prerequisite for depression diagnosis. ‘CDE’ and ‘GHI’ are superficial and deep symptom groups respectively. In the moderate symptoms network, Symptom ‘B’ dives to establish connect with ‘GHI’ and to replace Symptom ‘G’ in ‘GHI’ groups. In the severe symptoms network, ‘BHI’ forms an independent network, which shows that the participant may have symptom ‘H’ and ‘I’ if he has severe symptom ‘B’. This result implies that serious lack of interest (symptom ‘A’) and severe depression mood (symptom ‘B’) represent two types of severe depression, of which type B is likely to endanger the participant’s health.

4.3 Method Effect Analysis

Based on the fact of stratification, we can assume that there is only the same grade of symptom combinations between the depression symptoms, so the number of depression symptoms combinations for each grade is 2 ^ 9 – 1 = 511 species. We combine these 511 combinations with 11 methods and use the model to calculate the average effect of each method, of which the first 10 were behavior regulation and the last was clinical intervention. The results are shown in Table 5.

The results show that with the exacerbation of symptoms, both the behavior regulation and the clinical intervention can relieve the symptoms of depression. And the behavior regulation is more effective than clinical intervention for relieving mild

Table 5 The effectiveness of different methods on depression treatment

	Grade mild (%)	Grade moderate (%)	Grade severe (%)
Behavior regulation	18.3	59.7	91.7
Clinical intervention	1.0	80.0	98.0

depression symptoms. The clinical intervention for lessening moderate or severe depression symptoms has obvious advantages.

5 Conclusion

In this paper, we built an effective model to evaluate the effect of treatment methods for depression patients. The experiments show that the symptoms have unidirectional influence on the choice of regulatory methods. Our analysis shows that behavior regulation and clinical treatment have different advantages. This approach could be applied to more mental illnesses.

Acknowledgements This work was supported by Beijing Haola Technology Co., Ltd and the Mental Health Center of West China Hospital, Sichuan University. The online survey tool was designed by Beijing Haola Technology Co., Ltd. The authors acknowledge Prof. Zhang Jun at the Mental Health Center of West China Hospital, Sichuan University for his research assistance and insightful comments to the manuscript.

References

1. Anderson IM, Tomenson BM (1995) Treatment discontinuation with selective serotonin reuptake inhibitors compared with tricyclic antidepressants: a meta-analysis. BMJ 310:1433
2. Sansone RA, Sansone LA (2012) Antidepressant adherence: are patients taking their medications? Innov Clin Neurosci 9:41–46
3. Pampallona S, Bollini P, Tibaldi G, Kupelnick B, Munizza C (2004) Combined pharmacotherapy and psychological treatment for depression. Arch Gen Psychiat 61:714
4. Cuijpers P, Cristea IA, Ebert DD, Koot HM, Auerbach RP, Bruffaerts R, Kessler RC (2016) Psychological treatment of depression in college students: a metanalysis. Depress Anxiety 33:400–414
5. Joormann J, Gotlib IH (2017) Emotion regulation in depression: relation to cognitive inhibition. Cogn Emot 24:281–298
6. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT (2008) Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. PLoS Med. 5:0260–0268
7. Carl JR, Gallagher MW, Barlow, DH (2017) Development and preliminary evaluation of a positive emotion regulation augmentation module for anxiety and depression. Behav Therapy
8. Schuch FB, Vasconcelos-Moreno MP, Borowsky C, Zimmermann AB, Rocha NS, Fleck MP (2015) Exercise and severe major depression: effect on symptom severity and quality of life at discharge in an inpatient cohort. J Psychiat Res 61:25–32

9. Martin R, Dahlen ER. Cognitive emotion regulation in the prediction of depression, anxiety, stress, and anger
10. Millgram Y, Joormann J, Huppert JD, Tamir M (2015) Sad as a matter of choice? emotion-regulation goals in depression. Psychol Sci 26:1216–1228
11. Kroenke K, Spitzer RL (2002) The PHQ-9: a new depression diagnostic and severity measure. Psychiat Ann 32:509–515
12. Pearlstein TB, Zlotnick C, Battle CL, Stuart S, O'Hara MW, Price AB, Grause MA, Howard M (2006) Patient choice of treatment for postpartum depression: a pilot study. Arch Womens Ment Health 9:303–308
13. Mikolov T, Chen K, Corrado G, Dean J (2015) Efficient estimation of word representations in vector space. In: IJCAI international joint conference on artificial intelligence, Jan, pp 4069–4076 (2015)
14. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification

A Bayesian Network Approach for Discovering Variables Affecting Youth Depression

Euihyun Jung

Abstract Bayesian Networks have been used for data mining in many domains, but they have been rarely adopted in educational domain. In this paper, we model a Bayesian Network to discover which variables are in charge of youth depression and how strong the variables influence. For this study, Korean Children and Youth Panel Survey data are used and Markov Blanket is adopted to learn the Bayesian Network and to choose the relevant variables. In the results, “life satisfaction”, “social withdrawal”, “mobile phone dependency”, “attention”, “caregiver abuse”, and “aggressiveness” are extracted as the relevant variables to youth depression, therefore caregivers should pay attention to these variables of youths to reduce their depression. This paper shows Bayesian Networks are quite effective in finding the causal variables and their effects in educational domain.

Keywords Bayesian network · Markov blanket · Data mining · Depression

1 Introduction

From the point of appearance, Bayesian Networks (BNs) [1] have fascinated researchers of various domains such as Biology, Health, Finance, etc. [2–4]. A BN is a directed acyclic graph where the nodes represent variables and the edges show probabilistic dependencies between the nodes. Due to their compact and natural manner in describing probabilistic relations, BNs are useful not only for causal discovery, but also for prediction, classification, and diagnosis [5].

Given some dataset with BN learning algorithms, researchers can find a BN structure to show the causal semantics between the variables in the dataset. Once a BN structure is discovered from the data, then it can be used to reveal many meaningful aspects of the data. In a BN, every edge from a variable X to a variable Y means that X will cause Y probabilistically. Also, since every variable in a BN has a conditional

E. Jung (✉)
Department of Convergence Software, Anyang University, Anyang City, South Korea
e-mail: jung@anyang.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_43

probability and the probability can be changed, researchers can conduct various data mining by monitoring how the changes of variables' probabilities influence to other variables.

Although BNs have been intensely used in a lot of domains, they are rarely adopted in educational domain. For now, educational researchers have preferred traditional statistic methods because BNs are relatively new to them and require the considerable knowledge of probability. Also, the conventional statistical tools don't support BNs yet or the tools frequently require very expensive add-ons.

Since youth depression is considered a serious issue for youth mental condition, researchers in educational domain have tried to find the relevant variables of youth depression and have suggested the solutions to ease it [6, 7]. For this reason, a lot of studies have been conducted on the subject with various statistical methods, but BNs have not been virtually unexplored as a tool to aid educational researchers.

In this paper, we adopt BNs to discover which variables are directly and indirectly related to youth depression and how much the variables affect. For this, we preprocess the Korea Children and Youth Planet Survey (KCYPS) data [8] consisting of 40 variables. In order to find a BN structure and to select the relevant variables, the Incremental Association Markov Blanket (IAMB) [9] are used with the data. From the analysis with the BN structure, we conclude depression is affected by "life satisfaction", "social withdrawal", "mobile phone dependency", "attention", "caregiver abuse", and "aggressiveness".

2 Discovery of a Bayesian Network Structure

2.1 Overview of Dataset

This study used the data from the Korean Children and Youth Panel Survey (KCYPS) longitudinal data collected by the National Youth Policy Institute (NYPI) of South Korea [8]. The sample for the KCYPS was selected using stratified multi-stage clustering and was a nationally representative sample of Korean youth. The KCYPS conducted seven follow-up surveys from 2010 to 2016. At the time of the first study (2010), the children were in the 4th grade classes. For this study, data from the third wave (2012) were analyzed and the children were in the 6th grade classes of an elementary school.

Generally, data is frequently corrupted with missing values and noise, data preprocessing has become the important step to improve the quality of the data. Originally, each variable has a 4-point Likert scale (strongly yes, likely yes, likely no, and strongly no) and the values are mapped to a binary scale for learning BNs. In this paper, a total of 1,604 subjects were included after dropping out those which had missing fields. The used 40 variables are summarized in Table 1.

Table 1 The variables used for discovering a BN structure

Description (*Variable name*)
Gender (*gender*), Dad's education (*dad*), Health (*health*), Grade of language (*grlang*), Grade of Mathematics (*gramath*), Grade of English (*graengl*), Satisfaction of grade (*grasat*), Mastery goals (*mas*), Life of satisfaction (*life*), Management study time (*man*), Behavior control (*act*), Social withdrawal (*with*), Aggressiveness (*aggr*), Accomplishment (*accom*), Attention (*atten*), Depression (*depress*), Taunting (*mlaug*), Bullying (*mrej*), Assaulting (*mbeat*), Threating (*mthre*), Taunted (*damlaug*), Bullied (*damrej*), Assaulted (*dambeat*), Threatened (*damthre*), Care neglect (*caneg*), Sense of community (*commu*), Caregiver abuse (*caabu*), School act (*schact*), School rule (*schrul*), School friends (*schfri*), School teacher (*schtea*), Multi-culture attitudes (*multi*), Sense of local community (*local*), Mobile phone dependency (*phone*), Parents know my friends (*papeknow*), Parents meet my friends (*papemeet*), Parents like my friends (*papelike*), Date with the opposite gender (*date*), Enjoyment of Computer game (*comgame*), Fandom act (*fandom*)

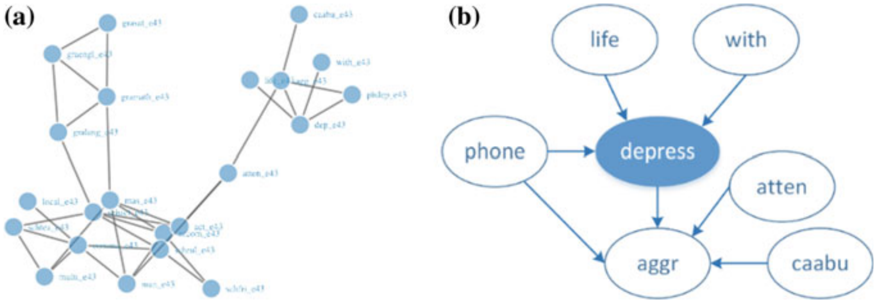


Fig. 1 **a** An initial discovered BN structure with the IAMB learning method. **b** The extracted variables related to "depression" with Markov blanket

2.2 The Discovered Bayesian Network Structure

In order to discover a BN structure, a statistical computing tool, R is used. We select IAMB [9] as a learning algorithm to find a BN structure and Fig. 1a shows the part of initial BN structure.

In order to extract the strongly related variables to depression, we find the Markov Blanket of *depress* variable. The Markov Blanket in a BN for node X_i which we denote by $MB(X_i)$ is a set of nodes composed of X_i's parents, its children and parents of its children. Formally the definition of Markov Blanket in a BN, or more general in a graph, is as follows.

$$MB(X_i) = Par(X_i) \cup Ch(X_i) \cup \bigcup_{Y \in Ch(X_i)} Par(Y) \tag{1}$$

Table 2 The strongly related variables to depression

Name	Description	Value
depress	Depression	1-yes, 2-no. Yes means I have it
life	Satisfaction of life	1-yes, 2-no. Yes means I'm satisfied with my life
with	Social withdrawal	1-yes, 2-no. Yes means I feel like it
aggr	Aggressiveness	1-yes, 2-no. Yes means I'm aggressive
caabu	Caregiver Abuse	1-yes, 2-no. Yes means I have experiences
phone	Mobile phone dependency	1-yes, 2-no. Yes means I'm dependent on it
atten	Attention	1-yes, 2-no. Yes means I have good attention

Using Markov Blanket, we determine the strongly related variables to depression from the discovered BN structure as shown in Fig. 1b. From the Markov Blanket of *depress* variable, *life*, *with*, *phone*, and *aggr* variables are directly related. Besides, *atten* and *caabu* variables are indirectly related. The variables of the Markov Blanket are summarized in Table 2.

3 Analysis

3.1 Interpretation of the BN Structure

Although the semantic relations among variables are revealed in the BN structure, it is not enough to explain how strong variables influence on depression and vice versa. Therefore, we adopt another tool, Netica [10], which shows the conditional probability distribution. Figure 2 shows how the initial BN structure with the probabilities is displayed in Netica. For example, in the *with* node, there is the value 26.8 which means P(yes = I feel like social withdrawal.) = 0.268.

3.2 Changing the Probability Values

Netica is good not only to show the probabilities of variables but also to enable researchers to monitor the degree of influences by changing the variables' probabilities.

Causal Reasoning. In BNs, it is possible to predict from causes to effects by changing the probabilities of causal variables. In Fig. 3, we maximize the "yes" value of *life* from 88.7 to 100.0%, the "no" value of *with* variable from 73.2 to 100.0%, and the "yes" value of *phone* variable from 76.6 to 100.0%. Then, the "yes" value of *depress* goes down from 5.98 to 1.0%. This is the lowest value of *depress*

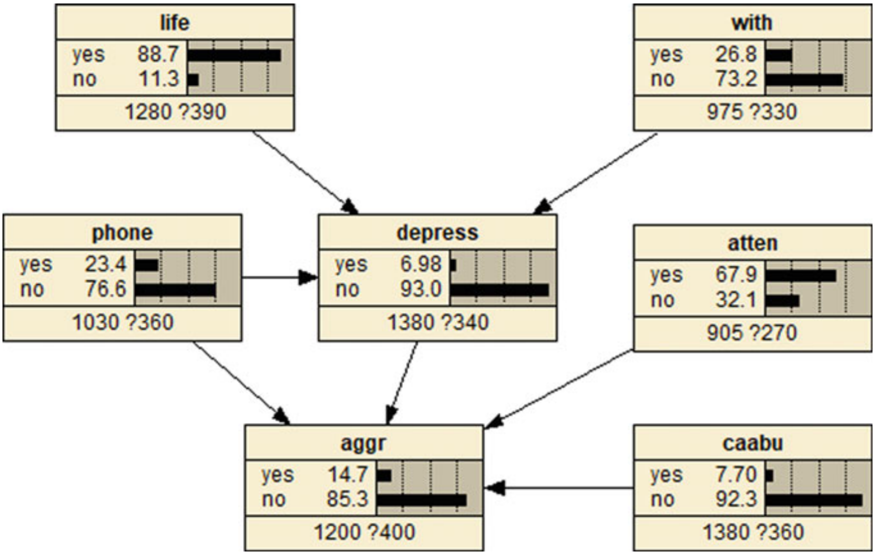


Fig. 2 The initial BN structure with the probabilities for variables

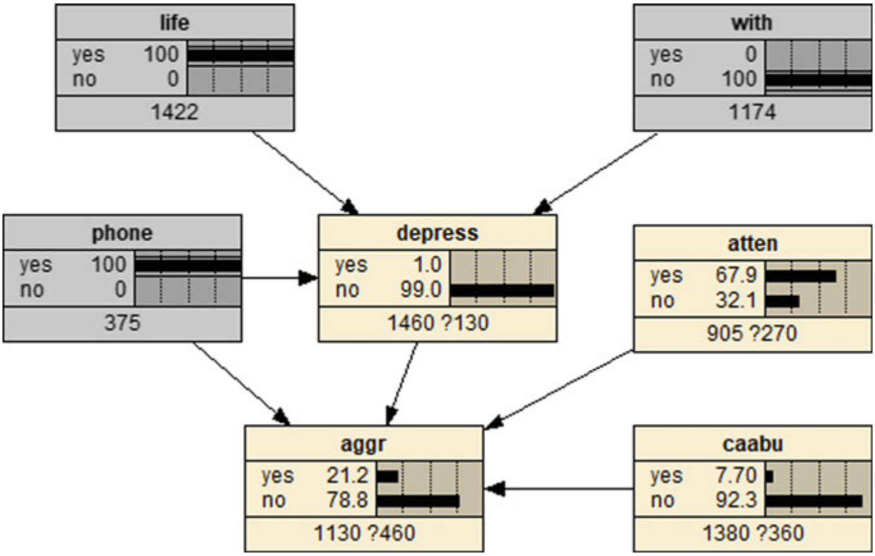


Fig. 3 We maximize the probabilities of the causal variables and observe the effect from the changes

and it is the case of the youth who is satisfied with his/her life, and doesn't feel like social withdrawal, and has phone dependency.

In the case shown in Fig. 3, the youth will not suffer from depression with a 99.0% chance, but when the variables have the opposite values, the value of depress

Table 3 Changes of causal variables' probabilities and their effects

Life	With	Phone	Depress (yes)	Depress (no)
Yes: 88.7, no: 11.3 (initial values)	Yes: 26.8, no: 73.2	Yes: 23.4, no: 76.6	6.98	93.0
Yes: 100.0	Yes: 100.0	Yes: 100.0	33.0	67.0
Yes: 100.0	Yes: 100.0	No: 100.0	4.4	95.6
Yes: 100.0	**No: 100.0**	**Yes: 100.0**	**1.0**	**99.0**
Yes: 100.0	No: 100.0	No: 100.0	1.4	98.6
No: 100.0	Yes: 100.0	Yes: 100.0	56.7	43.3
No: 100.0	Yes: 100.0	No: 100.0	58.0	42.0
No: 100.0	No: 100.0	Yes: 100.0	23.7	76.3
No: 100.0	No: 100.0	No: 100.0	20.3	79.7

rises to 58.0%. That is, if a youth is not satisfied with his/her life, feels like social withdrawal, and doesn't have phone dependency, the youth will be quite likely to suffer from depression.

This kind of experiment enables researchers to observe and compare various cases to determine how much a youth is likely to suffer from depression by changing the probabilities of the causal variables. In Table 3, we summarized the eight cases of maximizing the probabilities of the causal variables.

In order to figure out which causal variable has stronger effect than others, we conduct experiments by changing the value of a single causal variable while fixing the values of other two causal variables in the case of the lowest value of *depress*. First, we set *with* variable from "no" to "yes", then the value of *depress* rises from 1.0 to 33.30%. Second, when we change *life* variable from "yes" to "no", the value of *depress* is changed from 1.0 to 23.7%. Lastly, when we set *phone* variable from "yes" to "no", the value of *depress* is slightly changed from 1.0 to 1.4%. From the experiments, we conclude the influence of phone dependency is weaker than those of two other causal variables. In conclusion, in order to avoid youth depression, parents should keep their youths from getting social withdrawal and support them to be satisfied with their lives.

Evidential Reasoning. Another advantage of BNs is evidential reasoning. It is useful to explain the causal variables when the related evident is given.

In Fig. 4, we set the probability of *aggr* variable from 14.7 to 100.0% and this change makes *depress* variable be changed. During the investigation, we observe "yes" value of *depress* variable is increased from 6.98 to 21.7%. It is three times higher probability than before. It means if a youth is aggressive, the risk of having depression will be three times higher than a normal youth. Table 4 shows how the causal variable's probability is changed when the evidence is given.

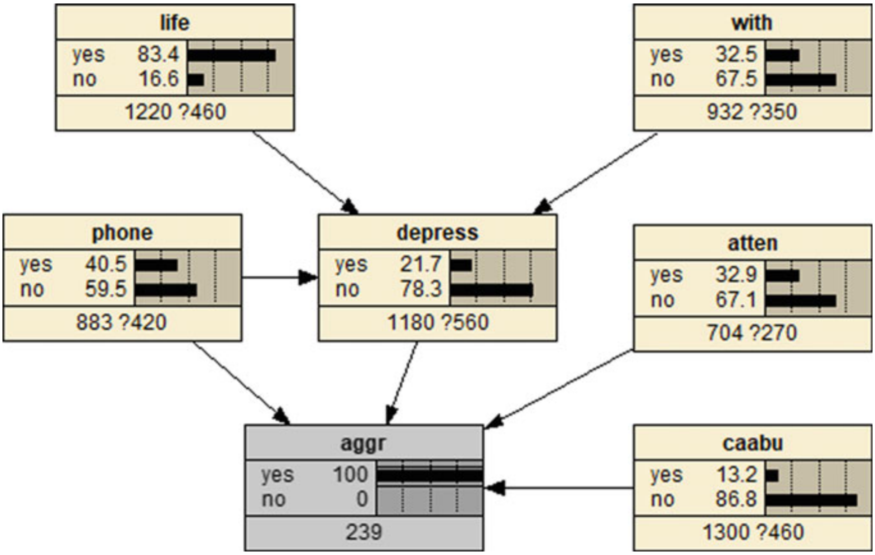


Fig. 4 When the evident is given, the BN structure can explain the causal variable

Table 4 Changes of *aggr* Variable's Value and Its Effects

Aggr	Depression (yes)	Depression (no)
Yes: 14.7, no: 85.3 (initial values)	6.98	93.0
Yes: 100.0	21.7	78.3
No: 100.0	4.44	95.6

4 Conclusion

Although BNs have been widely adopted in many multivariate domains, the usage in educational domain has been relatively low. However, like other domains, if BNs are adopted in the domain, it is expected BNs will reveal many useful insights that other research methods could not.

Currently, youth depression is becoming a serious social problem and parents' most concern. Therefore, the researchers in educational domain have conducted a lot of studies and reported meaningful results. However, in the previous research, it is difficult to predict which and how the multiple causal variables influence to youth depression.

In the paper, we try to find which variables are in charge of youth depression and how much the variables influence. We use the IAMB to discover a BN structure and apply Markov Blanket to extract "life satisfaction", "social withdrawal", "mobile phone dependency", "attention", "caregiver abuse", and "aggressiveness" as the strongly relevant variables to depression. From the causal reasoning, we conclude that social withdrawal and life satisfaction have much stronger effects on depression than phone dependency. Also, we observe aggressiveness is tightly related to depression from the evidential reasoning. In the future, we are going to examine the causal effects of other variables in detail and extend the research to find other meaningful features in educational domain.

References

1. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco, California
2. Lucas P (2004) Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care 10(5):399–403
3. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2006) Inference in Bayesian networks. Nature Biotechnol 24(1):51–54
4. Shenoy C, Shenoy PP (2000) Bayesian network models of portfolio risk and return. The MIT Press
5. Jensen FV (1996) An introduction to Bayesian networks. UCL press, London
6. Weisz JR, McCarty CA, Valeri SM (2006) Effects of psychotherapy for depression in children and adolescents: a meta-analysis. Psychol Bull 132(1):132–149
7. Craig WM (1998) The relationship among bullying, victimization, depression, anxiety, and aggression in elementary school children. Pers Individ Differ 24(1):123–130
8. Institute National Youth Policy (2010) The 2010 Korean children and youth panel survey I project report. Seoul, Korea
9. Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E (2003) Algorithms for large scale Markov blanket discovery. In: FLAIRS Conference, vol 2, pp 376–380
10. Norsys Software Corporaton: Netica is a trademarks of Norsys software Corporation, https://www.norsys.com/netica.html. Accessed 12 Feb 2018

Forecasting Macao GDP Using Different Artificial Neural Networks

Xu Yang, Zheqi Zhang, Laurie Cuthbert and Yapeng Wang

Abstract The objective of this paper is to forecast quarterly GDP in Macao using different neural network models. It is a challenge task due to the scarcity of determinant economic indicators and the scarcity of economic data. We compared the forecast errors of three different neural network models including Back Propagation (BP), Elman and Radial Basis Function (RBF). Elman has never been used in the GDP forecasting in literature, however in our results, Elman has the least forecasting error due to its recurrent network topology which can remember the history economic data.

Keywords Artificial neural network · Forecasting GDP · Back-propagation Elman · Radial basic function

1 Introduction

Forecasting Gross Domestic Product (GDP) is important to governments as part of their economic planning. One forecasting technique that can be used is the artificial neural network (ANN). ANNs can approximate general non-linear relationships which has been proved efficiency with good fault-tolerance and stability in many studies [1]. In recent years, it has been used to predict GDP in different countries. Some examples and used neural network models are summarized below.

X. Yang · Z. Zhang (✉)
School of Public Administrations, Macao Polytechnic Institute, Macao S.A.R, China
e-mail: zhangzheqitiny@yahoo.com

X. Yang
e-mail: xuyang@ipm.edu.mo

L. Cuthbert · Y. Wang
Information Systems Research Centre, Macao Polytechnic Institute, Macao S.A.R, China
e-mail: laurie.cuthbert@isrcmo.org

Y. Wang
e-mail: yapeng.wang@isrcmo.org

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_44

Back Propagation (BP) and Radial basis function (RBF) are the most popular neural network models used to predict GDP in different countries such as the U.S.A. [2], Indonesia [3], Turkey [4], Canada [5], Czech [6], European countries [7], Nigeria [8], and China [9] etc. The results demonstrated in these papers showed that by using different economic indicators such as inflation rate, interest rates etc. neural networks outperform Autoregressive Integrated Moving Average model (ARIMA models), linear models or other methods in terms of forecast accuracy.

Applying ANNs to the forecasting of GDP for a small territory, such as Macao, is a challenge task because of the scarcity of determinant economic indicators, and the scarcity of economic data.

1. *The scarcity of economic indicators*. Generally, two kinds of data (economic indexes and real time economic data) can be used to forecast GDP [3, 10]. Macao is a small economic entity where tourism and gambling form the dominant industry (63% in 2013 [11]); the success of this is very much dependent on the number of tourists and how much they are willing to spend. Therefore, it is very difficult to figure out the determinant economic factors.
2. *The scarcity of economic data*. According to the report from the Macao statistics office [11], the quarterly GDP of Macao consists mainly of 4 components: private consumption expenditure, government final consumption expenditure, investment and net exports goods and services (export minus import). These can be used as the inputs to forecast the next quarter, but this data (except the investment, which was missing) has been recorded only from the first quarter of 2001 to the second quarter of 2016 [11], so there are only 62 sets of data for training, many fewer than would normally be expected for training neural networks to provide accurate prediction.

In this paper, we investigate the performance of ANNs to forecast quarterly Macao GDP and compare the forecasting accuracy of three different NN models: BP neural network, RBF neural network, and Elman neural network. In literature, Elman neural network has never been used to forecast GDP. However, the results demonstrated in this paper show that Elman gets the best forecasting accuracy because of its recurrent neural network topology which can remember the history input.

2 Three Neural Network Models (BP, Elman, and RBF)

Artificial neural network (ANN) [12] is a logical computer programming technique which is based on operating mechanism of the way that brain solves problems.

ANN consists of multiple layers (input layer, hidden layers and output layer). Simple processing units called as nodes that are located in each layer. For each layer, the number of nodes can be different. Each node can interconnect to the other nodes in neighboring layers with weights which give the mathematical value of the relative power of information's connections. Input layer is grouped by the independent variables while output layer is identified by the dependent

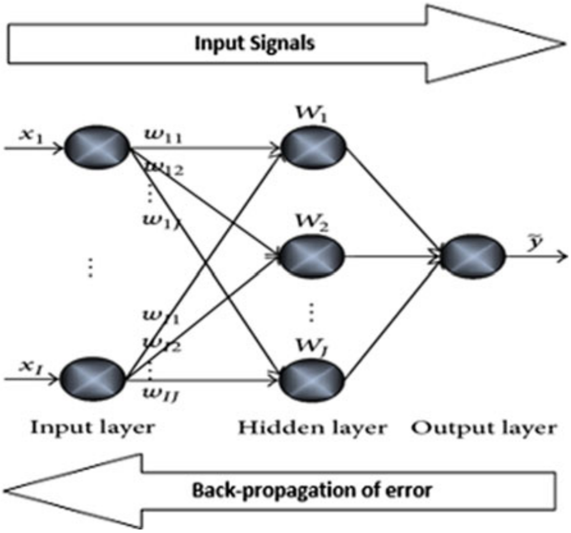


Fig. 1 Structure of back propagation neural network

variables. Transfer function is used for the conversing input to range of output in a neural network.

2.1 *Back Propagation Network*

BP network [13] is multi-layer feed forward neural network, which is one of the most widely used neural network model. The structure of BP network is shown in Fig. 1.

The training datasets are used for training the BP network, which consist of input data and the corresponding the target output data. The training datasets are fed to the network sequentially in an interactive manner, the weights of connections are corrected during the training of the networks to the desired behavior. This training iterations continues correcting the connections' weights until the desired mapping is reached. The backpropagation algorithm calculates the error contribution of each neuron after a batch of data is processed, then a Gradient Descent processing technique is performed to minimize the error function in weight space.

BP network is more sensitive to initial weights, it is not guaranteed to find the global minimum of the error function. Therefore, it is easy to reach a local minimum.

2.2 *Elman Network*

Elman neural network (ERNN) [14] employed by Jeff Elman is a kind of recurrent neural network. The basic structure of Elman network is showed in Fig. 2. In this

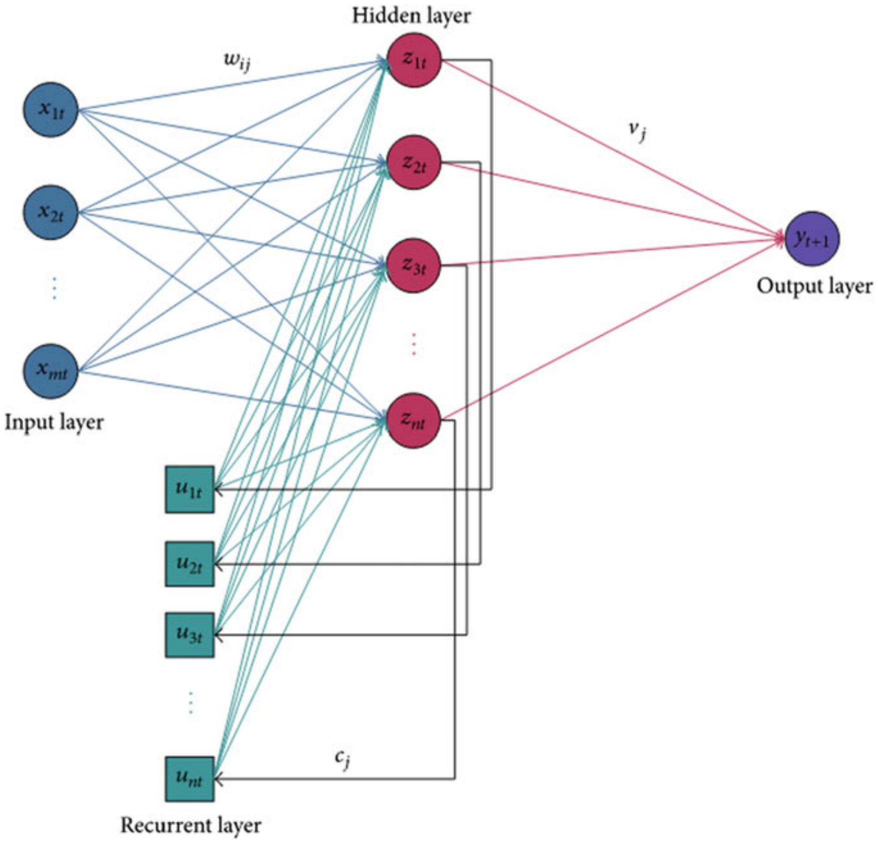


Fig. 2 Structure of Elman network [14]

network, the outputs of the hidden layer are fed back onto themselves through a buffer layer, named the recurrent layer. Through this way, ERNN can learn and generate temporal or spatial patterns. The hidden neurons are connected to the corresponding recurrent layer neurons by a constant weight 1. Therefore, the recurrent layer stores a copy of the state of the hidden layer by one step before. The number of neurons for the recurrent layer and the hidden layer shall be the same.

Elman neural can be used for speech processing, dynamic system modelling and control, time series prediction due to its sensitivity to the history of input data. It may not suitable for some time critical applications due to its converge speed [15].

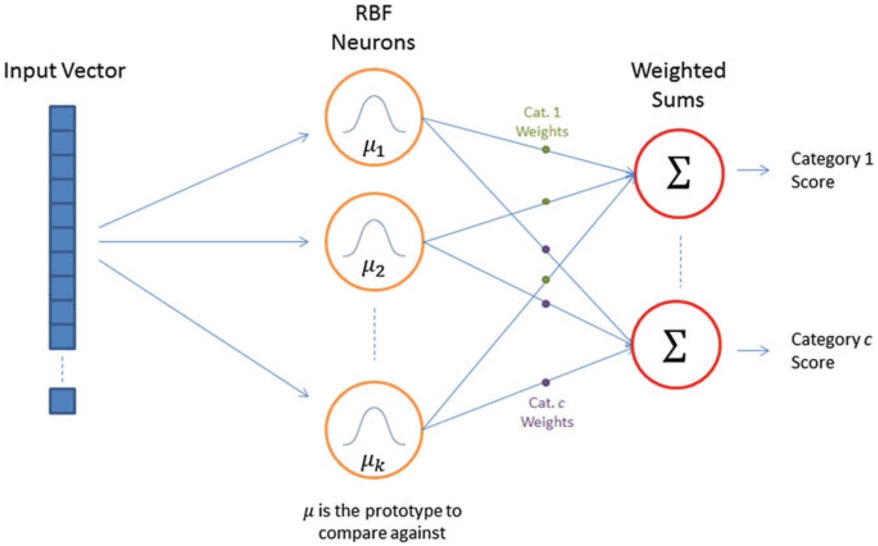


Fig. 3 Structure of RBF network [17]

2.3 Radial Basis Network

RBF network was first formulated in 1988 by Broomhead and Lowe [16]. It is a feed-forward neural network. RBF network typically has an input layer, a hidden layer with a non-linear radial basis functions as activation function and a linear output layer which is the combination of radial basis functions of the inputs and neuron parameters. Figure 3 shows the structure of RBF.

The radical basis function in hidden layer is radial symmetry of central point as well as attenuate non-negative, non-linear function. The function is different from other feed forward function which is overall response. It is partial response. The hidden layer transforms the lower dimensional input data into higher dimensional space. In output layer, it adjusts linear weight by linear optimization. RBF networks can tolerate more input noises than traditional ANNs (e.g. fully connected cascade networks) and therefore are more robust [18].

3 Data Collection and Normalization

3.1 Data Collection

The GDP of Macao mainly consists of 4 parts, private consumption expenditure, government final consumption expenditure, investment and net exports goods and services (export minus import). Figure 4 shows the four main components of GDP [11].

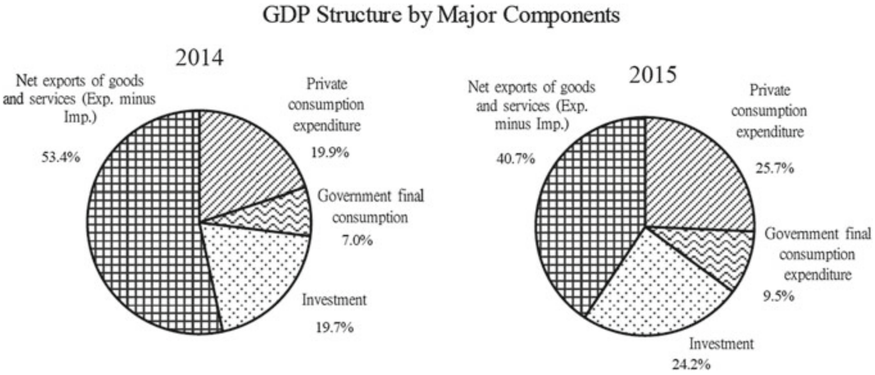


Fig. 4 Macao GDP structure

Totally 62 sets of data were collected from the first quarter of 2001 to the second quarter of 2016 except the data of investment which was missing. According to [11], gross capital formation is a component of the expenditure on GDP. It shows how much value is invested rather than consumed. Therefore, in this paper gross capital formation is used to replace the investment as one of the inputs to predict the quarterly GDP.

3.2 Data Normalization

Data normalization [19] is the preprocessing stage of design neural network. Normalization is used to transform each of the input and output data in the same range of values for the ANN model. It can guarantee stable convergence of weight and biases as well as improve convergence speed of algorithm and accuracy of network. There are two popular methods to perform data normalization: min-max normalization and zero-mean normalization.

In Min-max normalization, the raw data will be scaled to the range of data in [0, 1] or [−1, 1] by linear transformation. However, this method is not robust. It is highly sensitive to outliers. If a new data is inputted, maximum value or minimum value of the data set probably change. Then data need to be normalized again as shown in (1).

$$\mathrm{z} = \frac{x - \min(x)}{[\max(x) - \min(x)]} \tag{1}$$

Zero-mean normalization makes the values of each feature in the data have zero-mean and unit-variance. The arithmetic mean and standard deviation of the given data are calculated. The data, after normalization, follows normal distribution. However, this method may not guarantee a common numerical. Moreover, if the input data are not following Gaussian distribution, this method does not retain the distribution of the input data at the output. Zero-mean normalization is shown in (2).

$$x = (x - \overline{x})/\sigma$$
$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \, \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{2}$$

In this work, the input data and output quarterly Macao GDP are not Gaussian distributed. Therefore min-max normalization is more appropriate to solve the forecasting problem.

3.3 The Dataset of Training and Evaluation

To train NNs, the dataset should be divided into three distinct sets for training, testing and validation [14]. In this work due to the scarcity of economic data, the testing phase has to be omitted. The collected data are divided into 2 sets. The data collected from 1st quarter of 2001 to the 2nd quarter of 2014 which occupy approximate 90% of all data are used for training, and the rest of 10% of all data is used for validation.

3.4 Performance Evaluation

In this work, the forecast error between predicted output and actual output is evaluated by mean square error (MSE) and mean absolute error (MAE) on different NNs. The formula of MSE is shown in (3):

$$MSE = \frac{1}{N}\sum_{\mathrm{t}=1}^{\mathrm{n}} (\mathrm{PV_t} - \mathrm{RV_t})^2 \tag{3}$$

where n is the number of data. PV is the predicted output, and RV is the actual output. MSE is calculated in 10 billion MOPs.

The formula of MAE is shown in (4):

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \frac{ABS(PV_t - RV_t)}{RV_t} \tag{4}$$

4 Neural Network Topology Design

4.1 Input and Output Nodes of NN Models

Since Macao GDP are calculated by the sum of four different components, and the output is the predicted next quarterly Macao GDP. The problem becomes quite straightforward. In the different NN models there are five input nodes and one output node. Five inputs are four components of GDP (quarterly gross capital formation is used to replace investment) and plus previous quarterly Macao GDP. The output is the next quarterly Macao GDP.

4.2 Hidden Layers and Hidden Layer Nodes in BP and Elman

A major problem in designing a neural network is to determine the optimal number of hidden layers and number of neuros on each layer.

Normally increasing the number of hidden layers can improve the accuracy of network; on the other hand, it also increases the computation time and the chance of overfitting, i.e. the resulting neural networks overestimate the complexity of the training data [20]. However, a single layer is appropriate for most economic applications [19]. As we have a relatively small sample size and the number of parameters to be estimated, in this study only one hidden layer is used.

The hidden neurons can affect the output error of neural networks. Fewer hidden neurons can cause large training errors due to underfitting. On the other hand, the excessive hidden neurons can also lead to overfitting.

In last couple of decades, various criteria were proposed for fixing the number of hidden layer nodes. However, there is no a general formula for calculating the number of hidden layer nodes to guarantee the forecasting performance. Most researchers confirm the fixed number of hidden layer nodes by trail-and-error. The selection is based on the lowest errors. There are some experience formulas which were found and used by some researchers such as, $\sqrt{N+1}$, log2N, $\sqrt{NL}$, and $a + \sqrt{N+L}$, etc. [21]. N is the number of input, ‘L’ is the number of output and ‘a’ is the number from 0 to 10. By summarizing these experience formulas, a range of the number of hidden layer nodes was listed, which is from 1 to 13. Therefore, experiments were carried out to put 1 to 13 nodes in the hidden layer for both BP network and Elman network, shown in Table 1.

Therefore, in BP the best number of hidden layer nodes is 6, and in Elman is 8.

Table 1 The MSE performance for different number of hidden layer nodes

Nodes	1	2	3	4	5	6	7	8	9	10	11	12	13
BP (MSE)	0.623	0.12	0.116	0.084	0.089	**0.082**	0.092	0.101	0.137	0.114	0.122	0.189	0.145
ELM (MSE)	0.623	0.151	0.119	0.105	0.076	0.073	0.067	**0.061**	0.066	0.071	0.073	0.079	0.073

Table 2 Learning rate comparison

Leaning rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	Adaptive
BP (6 hidden nodes) MSE	0.144	0.263	0.145	0.108	0.131	0.121	0.18	0.163	0.085	0.108	**0.082**
ELM (8 hidden nodes) MSE	0.165	0.188	0.117	0.119	0.125	0.114	0.117	0.108	0.105	0.109	**0.061**

4.3 The Learning Rate in BP and Elman

Learning rate [13] is an important parameter in BP and Elman neural network. It is a proportionality parameter which defines the step length of each iteration. The value of the learning rate should be large enough to maintain a faster learning process but small enough to guarantee its effectiveness.

However, during training if learning output has a huge difference with the target, the high learning rate is more effective. On the contrary, when learning output is nearly close to target, a small learning rate is required. Adaptive learning rate can solve the problem. It adaptively changed the learning rate based on the step length of reaching target. The following Table 2 compares the forecast accuracy of adaptive learning rate and fixed learning rate which varies from 0.1 to 0.9. The results show that adaptive learning rate gets the best performance.

4.4 Topology Design in RBF

RBF [18] neural network has flowing tunable parameters:

- The type of radial function for the hidden units.
- The distance types.
- The number of radial functions
- The spread or radius of the radial functions.
- The center of the radial functions (location of the hidden units).

Gaussian function is normally used as the radial function and Euclidean distance as the distance type.

The technique of finding the suitable number of radial functions is very complex since excessively large networks are inefficient and sensitive to over-fitting with poor performance. In this work we used a simple incremental method to determine the

Table 3 Forecasting comparison

<table>
<tr><th></th><th>Hidden layer</th><th>Hidden nodes</th><th>Learning rate</th><th>Quarterly GDP (MSE)</th><th>Quarterly GDP (MAE)</th></tr>
<tr><td>BP</td><td>1</td><td>6</td><td>Adaptive</td><td>0.082</td><td>0.044</td></tr>
<tr><td>Elman</td><td>1</td><td>8</td><td>Adaptive</td><td>0.061</td><td>0.039</td></tr>
<tr><td></td><td>Radial functions</td><td>Spread</td><td colspan="3"></td></tr>
<tr><td>RBF</td><td>54</td><td>510</td><td colspan="2">0.206</td><td>0.053</td></tr>
</table>

number of RBFs, through experiments we found out that to solve the forecasting problem the appropriate number of radial function is 54.

The values chosen for the spread have great effects on the generalization abilities of the network. Spread is the extension speed of radial basis function which is constant for the radial basis layer. The higher spread, the more smoothness of the simulation result. However, the higher spread also can lead to complex calculation and less sensitivity. In this work, the spread of radial functions of the hidden units are set to a fixed value obtained through experiments. In the experiments, the value of spread changes from 1 to 1000. When the spread is 510, it gets the least forecasting error.

5 Result Comparison

In this paper, we compared the forecasting performance with three different neural network models: BP, Elman and RBF. The collected dataset only contains 62 set of data from the first quarter of 2001 to the second quarter of 2016. 90% of the data were used for training, 10% of data were used to validate the forecast accuracy. The tuning parameters used in each neural network were adjusted through experiments, and the results are listed in Table 3. Figure 5 demonstrates the whole training process and compares the performance of different neural network models with the actual quarterly GDP in every quarter.

All the different types of neural networks could achieve the accepted forecasting accuracy. However, Elman outperforms BP and ELM network models, gets the lowest forecasting errors. This is due to Elman is a kind of recurrent neural network which can remember the history input.

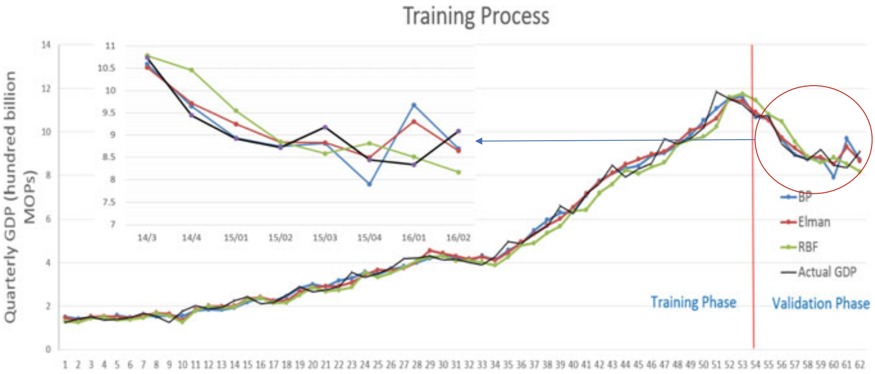


Fig. 5 Forecasting result comparison

6 Conclusion

In this paper, we have applied three different neural network models (BP, Elman and RBF) to forecast quarterly GDP in Macao. In literature, BP and Elman are the most popular neural network models used to forecast GDP. No paper compares the forecast performance of Elman.

The forecasting system used the previous quarterly four main components of Macao GDP to estimate the next quarterly Macau GDP without involving any other economic indicators. Although the datasets collected for this study is quite small, only about 14 years of quarterly data were collected for training, additionally the data for one important GDP component (investment) was missing, gross capital formation was used to replace investment as one of the input. The experiments results showed that Elman got the lowest forecasting error (MAE is 0.0385) because of its recurrent network topology which can remember the history of economic data. Also, the results demonstrated that without any economic indicators, only using GDP components to forecast future GDP can get relative accurate forecasting in small territory like Macao.

References

1. Haykin S Neural networks and learning M
2. Islam R (2013) Predicting recessions: forecasting US GDP growth through supervised learning. Stanford Univ, Dep. Electr. Eng
3. Liliana, Napitupulu TA (2012) Artificial neural network application in gross domestic product forecasting an Indonesia case. J Theor Appl Inf Technol 45, 410–415
4. Karaatli M (2012) Using artificial neural networks to forecast GDP for Turkey. In: 3rd international symposium on sustainable development. Sarajevo
5. Tkacz G (2001) Neural network forecasting of Canadian GDP growth. Int J Forecast 17:57–69
6. Dvořáková L (2017) CZ GDP prediction via neural networks and box-jenkins method. In: Innovative economic symposium

7. Vrbka J (2016) Predicting future GDP development by means of artificial intelligence. Littera Scr. 9:154–167
8. Maliki OS, Emmanuel I, Obinwanne EE (2014) Neural network applications in the forecasting of GDP of Nigeria as a function of key stock market indicators. Adv Appl Sci Res, 204–212
9. Zhao J, Wang X, Wu Z (2008) Forecasting GDP growth based on ant colony clustering algorithm and RBF neural network. In: 2008 IEEE international conference on automation and logistics, pp 1839–1843. IEEE
10. Kadirkamanathan V, Niranjan M, Fallside F (1991) Sequential adaptation of radial basis function neural networks and its application to time-series prediction. In: Advances in neural information processing systems, pp 721–727
11. DSEC of Macao: DSEC database, http://www.dsec.gov.mo/TimeSeriesDatabase.aspx. Accessed 10 Feb 2018
12. Samarasinghe S (2016) Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. CRC Press
13. Rojas R (1996) Neural networks: a systematic introduction. In: Neural networks, vol 502
14. Wang J, Wang J, Fang W, Niu H (2016) Financial time series prediction using elman recurrent random neural networks. Comput Intell Neurosci
15. Hush DR, Horne BG (1993) Progress in supervised neural networks. IEEE Signal Process Mag 10:8–39
16. Broomhead DS, Lowe D (1988) Multivariable functional interpolation and adaptive networks. Complex Syst. 2:321–355
17. McCormick C Radial basis function network (RBFN) tutorial. http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/
18. Esmaeili A, Mozayani N (2009) Adjusting the parameters of radial basis function networks using particle swarm optimization. In: 2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, CIMSA 2009, pp 179–181
19. Schumacher C, Breitung J (2008) Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. Int J Forecast 24:386–398
20. Jinchuan K, Xinzhe L (2008) Empirical analysis of optimal hidden neurons in neural network modeling for stock prediction. In: Proceedings—2008 Pacific-Asia workshop on computational intelligence and industrial application, PACIIA 2008, pp 828–832
21. Levich RM, Thomas LR (1993) The significance of technical trading-rule profits in the foreign exchange market: a bootstrap approach. J Int Money Financ 12:451–474

Edge Detection and Symmetry Method for Cleft Lip and Palate Children Using Image Processing

Nutthisara Choolikhit, Wararat Songpan, Monlica Wattana and Ngamnij Arch-in

Abstract To evaluate the treatment result of the cleft lip and palate children, it needed the specialists to evaluate it on the basis of the nasolabial method. The vital principle of this method was the result of operation of the lip shape to see whether it was symmetrical after being operated. The problem was that it needed at least five specialists to evaluate the treatment result through the nasolabial method. In case there was any disagreement between the specialists, it needed to hold the meeting to summarize the result of the lip operation. According to the found problem, this study proposed the suitable method for the edge detection, and symmetry for the cleft lip and palate children using the image processing based on the criteria of the shape of the vermilion border. For the procedures of analysis, the 5 tested images of the child patient who has been operated to modify the cleft lip and palate condition were examined to measure the efficiency. The comparison of the methods to finding out the suitable edge detection was carried out by Roberts, Prewitt, Sobel, Canny, and Laplacian of Gaussian (LoG) methods. The results of the study revealed that the average of LoG was the suitable method of finding out the shape of the vermilion border as compared to other methods. It had the least pixel difference of all as compared to the actual edge image. Moreover, it was found that the left lip and the right lip had the least distance pixel difference when finding out the symmetry by clipping the image. Thus, this proposed method could be the guideline of evaluating the treatment more accurately and efficiently.

N. Choolikhit (✉) · W. Songpan · M. Wattana · N. Arch-in
Faculty of Science, Department of Computer Science, Khon Kaen University, Khon Kaen, Thailand
e-mail: nuttcho@kku.ac.th

W. Songpan
e-mail: wararat@kku.ac.th

M. Wattana
e-mail: monlwa@kku.ac.th

N. Arch-in
e-mail: ngamnij@kku.ac.th

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_45

Keywords Cleft lip and palate children · Image processing · Edge detection
Nasobial method · Vermilion border

1 Introduction

Nowadays, the image processing took more important role in the medicine. Vezzetti et al. [1] had utilized the image processing to automatically diagnose and formalize prenatal cleft lip with representative key points and identify the type of defect (unilateral, bilateral, right, or left) in three dimensional ultrasonography (3D US). In addition, Tonpho [2] had implemented the image processing to analyze the lung abnormality by analyzing the chest x-ray. The experimental results revealed that the overall accuracy was at approximately 95%. As the above-mentioned studies, it could be seen that the image processing could be applied to the medical affairs efficiently.

The cleft lip and palate was one of the facial abnormalities occurring since the children were born. It was usually found on the upper lip, the nose, the upper gum, and the palate. It could be noticed that the upper lip, the upper gum, and the palate were cleft. Although there was not the cleft on the nose, it appeared the deformation. The incidence of Thailand was approximately 1–1.5:1000 of the survived newborn infants. As considered the data base of the population in 2012, it might be estimated that the incidence of disease in Thailand about the cleft lip and palate was approximately at least 800–1200 patients per year [3]. The procedures of the cleft lip and palate treatment were rather complicated since it concerned the abnormalities in different parts of the organ on the face such as the lip, the nose, teeth, the gum, the palate, and the maxillary bone. Thus, the treatment needed to be carried out by the specialists from different study fields comprising the aesthetic plastic surgeon, the otolaryngologist, the orthodontic dentist, and the speech-language pathologist. The cleft lip and palate needed to be treated by means of surgery and treated continuously since being born till the organs on the face fully grown, which was aged 20 years. The organs of the face tended to grow abnormally. When the patient was treated by surgery to modify the cleft lip and palate, it needed to evaluate the treatment result by analyzing the image of the patient after the surgery. The specialists then evaluated the treatment result by rating the score regarding their own opinions without the outstanding criteria of the evaluation. The obtained evaluation resulted in the error and inaccuracy. As a result, it needed to hold the meeting to come up with the consensus.

According to the mentioned problem, this present study attempted to investigate the method of evaluating the edge detection and the lip symmetry after the surgery of the children patients who suffered the cleft lip and palate through the image processing. It was also to obtain the standardized criteria of the evaluation more outstandingly.

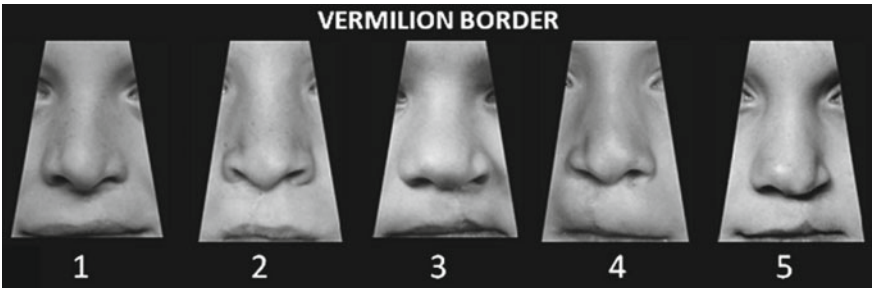


Fig. 1 Reference photographs for rating regarding the shape of the vermilion border of the nasolabial appearance. The numbers 1–5 referred to the scores from 1 representing a very good appearance, to 5 representing a very poor appearance [4]

2 Related Work

2.1 Nasolabial Appearance [4]

This research study adduced to the treatment evaluation of the patients with a popular method for nasolabial rating in the cleft lip and palate by modifying the Asher-McDade method. The nasolabial rating system for the cleft lip and palate was used for scoring the nasolabial appearance. In this index, four components of the nasolabial area were scored separately:

1. Nasal form (frontal view),
2. Deviation of the nose (frontal view),
3. Shape of the vermilion border, and
4. Nasal profile including upper lip.

This study investigated the model of finding out the edge detection to evaluate the lip symmetry regarding the item 3 (shape of the vermilion border) of the nasolabial appearance (Fig. 1).

Nalin Panthavong et al. [5] had explored the result of the cleft lip and palate patients, especially on the lip and nose, and the life quality after surgery in Laos PDR. Five evaluators consisting of the aesthetic plastic surgeon, the speech-language pathologist, the nurse, the dentist, and the outsider who was not related to the medicine evaluated the surgery result according to the criteria of the nasolabial appearance. Moreover, Mercado et al. [6] had studied about developing a yardstick of reference photographs for the nasoloabial appearance assessment of 5- to 7-year-old patients with the complete unilateral cleft lip and palate (CUCLP) by using the Asher-McDade method. It could be noticed that the evaluation criteria of the nasolablial appearance by the Asher-McDade method in the two above-mentioned studies was reliable and able to apply to evaluate the surgery result of the cleft lip and palate patients.

2.2 Edge Detection

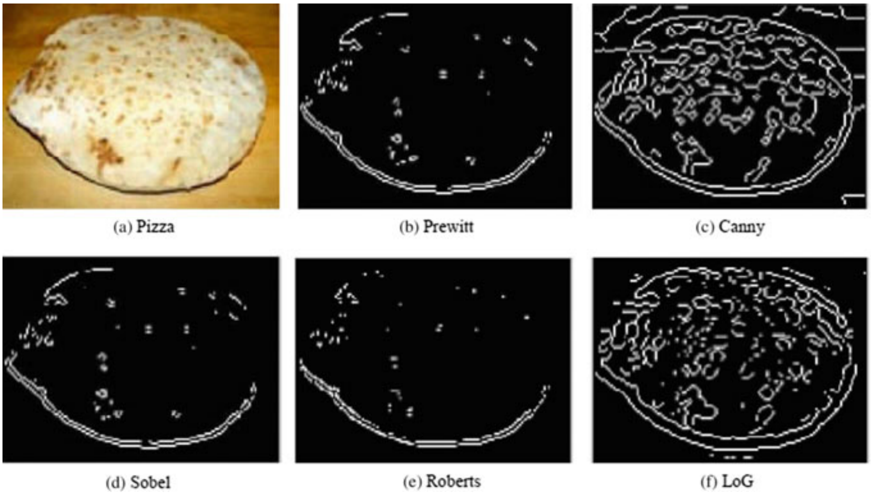
The edge detection was a basic tool used in the image processing, basically for the feature detection and extraction, which aimed to identify points in the digital image where brightness of the image changed sharply and found discontinuities. There were various methods of the edge detection. However, it could be categorized into two main methods; Gradient method and Laplacian method detailed as the followings.

1. Gradient method was the method of the edge detection by finding the highest and lowest points through the first derivative of the image. The edge would be counted for over the threshold, which encouraged the edge with the thick border. The examples of the methods of the edge defection included Roberts, Prewitt, Sobel, and Canny.
2. Laplacian method was another method of the edge detection consisting of Laplacian of Gaussian and Marrs-Hildret.

Acharjya et al. [7] had studied the efficiency comparison of the different methods of the image edge detection by comparing five methods of the edge detection consisting of sobel, Prewitt, Zero crossing, LoG (Laplacian of Gaussian), and Canny. According to this study, it could be observed that the performance of the Canny edge detection operator was much better than Sobel, Roberts, Prewitt, Zero crossing and LoG (Laplacian of Gaussian) in respect to the image appearance and the object boundary localization. In addition, Muthukrishnan and Radha [8] had explored the performance of various edge detection techniques for the image segmentation and also the comparison of these techniques such as Roberts edge detector, Sobel edge detector, Prewitt edge detector, Kirsch, Robinson, Marr-Hildreth edge detector, LoG edge detector and Canny edge detector. These detectors were carried out with the experiment by using the MATLAB software. The results revealed that Marr-Hildreth, LoG and Canny edge detectors produced almost the same edge map. The result of the Canny edge detector result was superior as compared to all edge detectors for the selected image since the different edge detections worked better under the different conditions (Fig. 2).

3 Proposed Methodology

There were totally six procedures for the proposed method for converting the pictorial information of the cleft lip and palate patient. Then, find the suitable method of the mouth detection and analyze the symmetry of the image. Each procedure was implemented the MATLAB 8.3.0.532 programming language, and performed on the computer equipped with a CPU of Intel, core i7@3.4 GHz, running Windows 10 Pro 64-bit and 8 GB memory. The procedures were detailed as Fig. 3.



(a) Pizza (b) Prewitt (c) Canny

(d) Sobel (e) Roberts (f) LoG

Fig. 2 Examples of the image edge detection through the edge detectors [9] **a** original image; **b** prewitt; **c** canny; **d** sobel; **e** roberts; **f** Laplacian of Gaussian

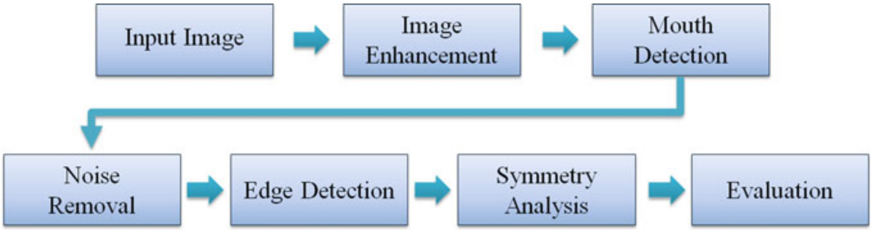


Fig. 3 Overview of the proposed methodology

3.1 Input Image

The data used for this research study were the images of the cleft lip and palate patient treated at the Center of Tawanchai, Srinagarind Hospital, Khon Kaen and convalesced from the surgery till the wound was fully recovered. The patient was appointed to follow up the surgery result. The image of the patient was also taken by the digital camera. The input image was in the RGB mode.

3.2 Image Enhancement

Since the input image was taken by the digital camera and it was in the RGB mode, the obtained image had the different colors depending on the lighting conditions. Thus, the color of the input image needed to be adjusted to be more complete and suitable for applying to the processing in the next procedure as shown in the followings.

3.2.1 Color Adjustment

It could be done by increasing the contrast, reducing the brightness, and increasing the depth of the green shade.

3.2.2 Changing RGB to YCbCr

It was the procedure of converting the input image from RGB to YCbCr so that it could better capture the complexion. This procedure could detect the position of the mouth more inclusively and accurately.

3.2.3 Changing the Image to the Gray Scale

This procedure encouraged the image to be easier and more convenient for the processing since each color image consisted of three images namely red, green, and blue tones. Thus, to access the image and the processing, it needed to access all the three data (red, green, and blue). However, the grayscale was easier and quicker because it was the single tone. That was, the brightness gradient was arranged from the dark to the white tones. The differences of brightness included 256 levels from 0 to 255 (Fig. 4).

3.3 Mouth Detection

The face image that had been already improved the quality was taken into account the processing in the next procedure. Firstly, capture the mouth position by implementing the CascadeObjectDetector [10]. Secondly, find the position of the left and right mouth corners. Thirdly, rotate and crop the image with a focus on the upper lip which was the area to be analyzed in the next procedure (Fig. 5).

3.4 Noise Removal

Firstly, the grayscale image was converted to the binary by specifying the threshold between 100 and 135. If the pixel was over the threshold, it would be represented with 0. However, it would be represented with 1 if the pixel was less or equal to the threshold. Secondly, the noise was removed by utilizing the imfill with the holes

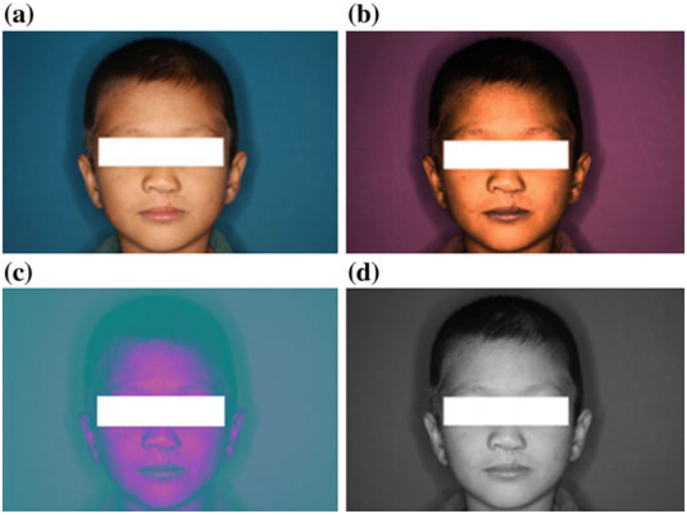


Fig. 4 **a** Input image; **b** contrast, brightness and green channel; **c** YCbCr image; **d** gray image

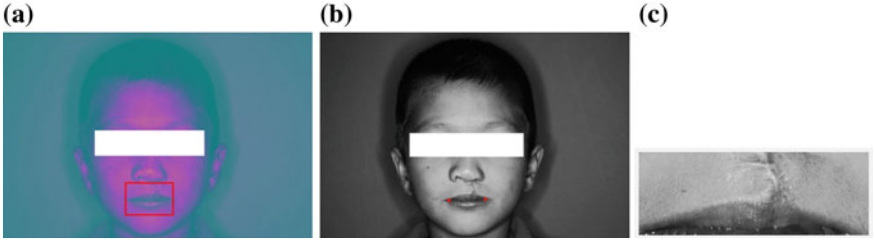


Fig. 5 **a** Detected mouth; **b** detected mouth corner; **c** rotated and cropped upper lip

parameter. Thirdly, the image was reversed from black to white and white to black since the needed object area of the image obtained from the threshold was black. The MATLAB program specified the black as the background whereas the object was the white. Thus, it needed to reverse the black image to the white one so that this program could process particularly the interesting part through utilizing the imcomplement function. Finally, the pixel link was categorized by utilizing the imclose function with the line parameter. Additionally, the noise was once removed by utilizing the imfill function with the holes parameter (Fig. 6).

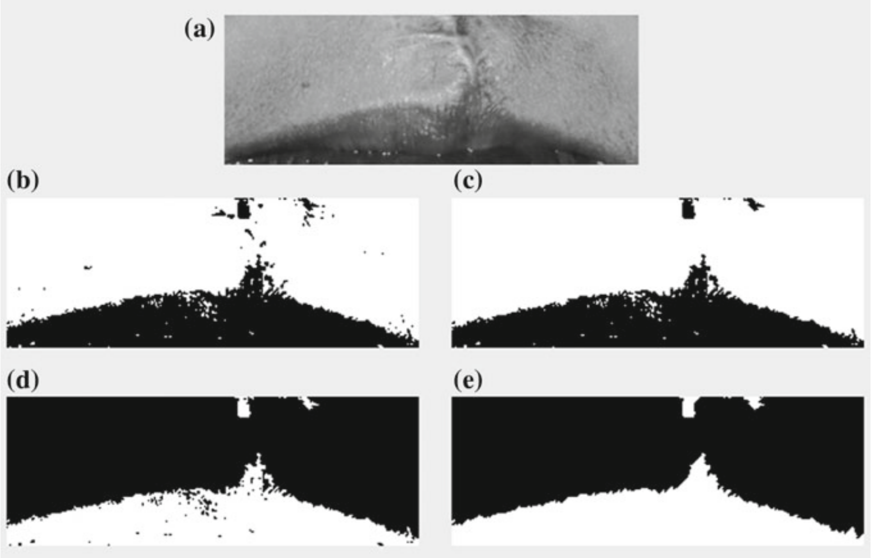


Fig. 6 **a** Gray image; **b** coverting to binary; **c** noise removal by utilizing the imfill; **d** converting the black image to the white image; **e** categorizing the pixel link

3.5 Edge Detection

The researcher investigated the method of the edge detection that was the most suitable for this present study by comparing all five methods consisting of Roberts, Prewitt, Sobel, Canny, and Laplacian of Gaussian (LoG) (Fig. 7).

3.6 Symmetry Analysis

The left and the right mouth edges were analyzed to see what level the symmetry was by flipping the image to compare the symmetry between the left and right lips through the flipdim function below (Fig. 8).

$$\text{flipdim(Img, 2);} \tag{1}$$

3.7 Evaluation

The imshowpair function with the diff parameter in the MATLAB program was utilized to compare the image before and after flipping. It provided the diff area

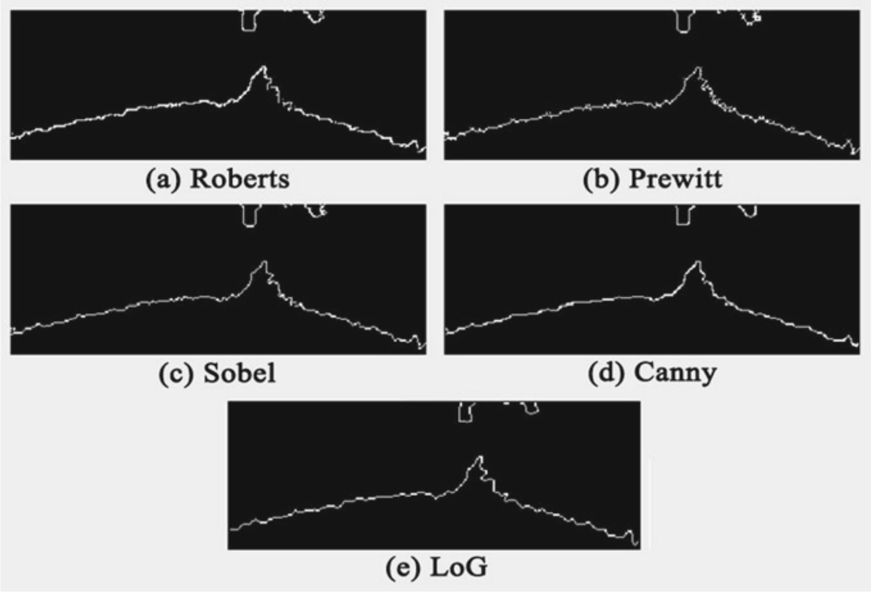


Fig. 7 **a** Roberts; **b** prewitt; **c** sobel; **d** canny; **e** Laplacian of Gaussian

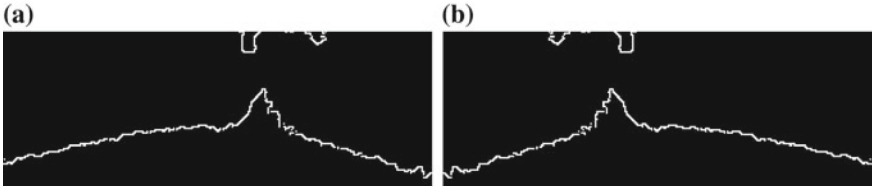


Fig. 8 **a** The image before flipping; **b** the image after flipping

to examine the lip symmetry. The researcher divided the evaluation into two parts. Firstly part, the pixel difference (pixel diff) and distance difference (distance diff) were measured for the methods of edge detection between actual edge of lip and the methods of edge detection. Secondly, the pixel diff and distance diff were measure for the symmetry image via methods of edge detection and flip images. The calculation of pixel diff and distance diff was as following,

1. The pixel diff was the count of the number of the white pixels (equal to 1) which measured actual edge of lips and detected lips of image as Table 1. In addition, flipping image after detected edge of lip as Table 2 was compared by using the function below.

$$\text{PixelDiffSym} = \text{sum(ImgDiff(:))}; \qquad (2)$$

Table 1 Comparison of the pixel diff and the distance diff of each edge detection method

Methods	Pixel diff						Distance diff					
	Img1	Img2	Img3	Img4	Img5	Avg.	Img1	Img2	Img3	Img4	Img5	Avg.
Roberts	913	887	508	446	673	*685.4*	824	1071	736	488	914	*806.6*
Prewitt	881	814	421	435	521	*614.4*	796	1061	600	501	803	*752.2*
Sobel	840	791	414	409	525	*595.8*	748	1023	590	483	722	*713.2*
Canny	831	832	443	**399**	577	*616.4*	753	1016	626	**453**	746	*718.8*
LoG	**830**	**782**	**406**	424	**489**	***586.2***	**739**	**952**	**553**	510	**712**	***693.2***

Table 2 Comparison of the pixel diff and the distance diff of symmetry through each edge detection method

Methods	Pixel diff						Distance diff					
	Img1	Img2	Img3	Img4	Img5	Avg.	Img1	Img2	Img3	Img4	Img5	Avg.
Roberts	605.5	523.5	323	324	481	*451.4*	902	693.5	488	412	2337	*966.5*
Prewitt	570.5	472.5	257	304	331	*387*	913	710.5	483	418	2278	*960.5*
Sobel	518.5	445.5	246	278	335	*364.6*	865	693.5	488	416	2318	*956.1*
Canny	542.5	480.5	283	270	403	*395.8*	872	659.5	498	**364**	2304	*939.5*
LoG	**499.5**	**430.5**	**244**	**253**	**303**	***346***	**864**	**628.5**	**480**	378	**2275**	***925.1***

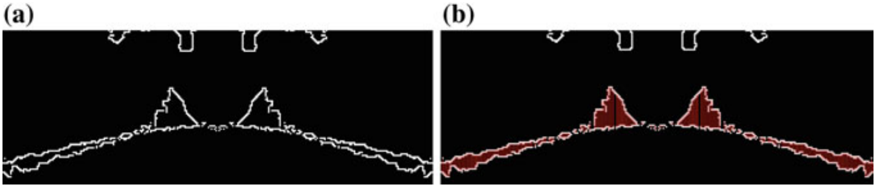


Fig. 9 **a** The pixel diff value; **b** the distance value between the left and right lips

2. The distance diff was the calculation of the distance both of compared actual-detected edge of lips and detected edge flipping of the left and right lip symmetry by whirling the loop to check the distance from the starting point to the ending point of each row. The distance of each row was combined as the whole distance. The example of image was shown distance diff area as Fig. 9b.

4 Experimentation and Result

According to the primary experiment, it was found that converting the RGB image to the YCbCr model could find out the position of the mouth more accurately. As for the mouth defection, the researcher compared five methods of the edge detection to find the suitable method for this present study, namely Roberts, Prewitt, Sobel, Canny, and Laplacian of Gaussian. The image of the edge detection through the different methods was compared to the actual lip edge of the original image. Then, the imshowpair function with the diff parameter was utilized to demonstrate particularly the diff area. After that, the image was calculated to find out the pixel diff and the distance diff of each method as shown in Table 1.

Moreover, the researcher used the image obtained from the five methods of the edge defection to find out the symmetry. The results of the pixel diff and the distance diff were shown in Table 2. The results of the edge detection through LoG had the least pixel diff and the distance diff. It meant that the obtained edge was the most approached to the actual mouth edge.

5 Conclusion

Regarding the proposed method for the edge detection of the child patient who was treated the cleft lip and palate through the surgery by comparing the efficiency of the suitable method, it was revealed that LoG had the value approached to the actual mouth edge of the patient. So, it was analyzed to find out the symmetry of the mouth shape based on the criteria of evaluating the shape of the vermilion border of the nasolabial rating in cleft lip and palate by modifying the Asher-McDade method.

The procedure of the edge detection was considered the vital procedure for analyzing the symmetry of the mouth shape more precisely and accurately before finding the differences between the left and right lips. Thus, this present study inputted the actual image to present to the specialists so that they could see the image more clearly. They could also see the symmetry of the actual mouth edge and could evaluate the image precisely and accurately. As a result, the evaluation result after the surgery of the child patient who was treated the cleft lip and palate came up with the efficient evaluation. For the future study, it should concern about finding the suitable method of image enhancing before presenting since the patient image was taken in the unequal lighting conditions. This result depends on the processing of the image adjustment.

References

1. Vezzetti1 E, Speranza D, Marcolin F, Fracastoro G (2016) Diagnosing cleft lip pathology in 3d ultrasound: a landmarking-based approach. Image Anal Stereol 35(1):53–65
2. Investigation of chest x-ray images based on medical knowledge and images processing. intelligent signal processing and communication systems (ISPACS), pp 978–981 (2010)
3. Khwanngern K (2014) The treatment evaluation of cleft lip and palate patient. Department of Surgery, Faculty of Medicine, Chiang Mai University, Thailand
4. Kuijper-Jagtman AM, Gunvor S (2009) Outcome of patients with cleft lip and cleft palate-operated at Mahosoth, Mitthaphab and Setthathirath Hospitals in Lao People's Democratic. J Craniofac Surg 20(1):1683–1686
5. Panthavong N, Pradabwong S, Luvira V, Khansoulivong K, Chowcheun B (2013) Outcome of patients with cleft lip and cleft palateoperated at Mahosoth, Mitthaphab and Setthathirath Hospitals in Lao People's democratic. J Med Assoc Thai 96(1):98–106
6. Mercado AM, Russell KA, Daskalogiannakis J, Hathaway RR, Semb G, Ozawa T, Smith A, Lin AY, Long RE (2016) The Americleft project: a proposed expanded nasolabial appearance yardstick for 5- to 7-year-old patients with complete unilateral cleft lip and palate (CUCLP). Cleft Palate-Craniof J 53(1):30–37
7. Acharjya PP, Das R, Ghoshal D (2012) Study and comparison of different edge detectors for image segmentation. Global J Comput Sci Technol Graph Vis 12(13):28–32
8. Muthukrishnan R, Radha M (2011) Edge detection techniques for image segmentation. Int J Comput Sci Inf Technol (IJCSIT) 3(6):259–267
9. Narendra VG, Hareesh KS (2011) Study and comparison of various image edge detection techniques used in quality inspection and evaluation of agricultural and food products by computer vision. Int J Agric Biol Eng 4(2):83–90
10. Elena A, Corneliu L (2015) A practical implementation of face detection by using matlab cascade object detector. In: International conference on system theory, control and computing (ICSTCC), vol 19, pp 785–790

An Effective Envelope Analysis Using Gaussian Windows for Evaluation of Fault Severity in Bearing

Nguyen Ngoc Hung, Jaeyoung Kim and Jong-Myon Kim

Abstract This paper proposes an efficient method for evaluation of the fault severity in bearing using the discrete wavelet packet transform (DWPT) and the envelope analysis. The acoustic emission (AE) signals for each defect are first decomposed to the sub-band signals. The envelope power spectrum analysis is performed on each sub-band to detect the frequency periodic impulses showing the abnormal symptoms of bearing defects. It is essential to select an optimal sub-band for reliable assessment of the fault severity in bearing. A ratio of defect spectral component to residual spectral component (RDR) is calculated from their envelope power spectrum using the Gaussian window for an optimal sub-band selection which shows clearly information about failures. As a result, the severe degree of bearing defects is assessed based on the RDR calculation. The effectiveness of the proposed scheme is validated through experimental results of evaluating the different fault conditions under variable crack size in bearing.

Keywords AE signal · Bearing fault severity · Envelope analysis · DWPT
Sub-band analysis

1 Introduction

Induction motors have been widely used in the industrial applications. They can often face unanticipated faults that can make unexpected interruptions, economic losses in production. Hence, the accurate evaluation and identification of several faults at the early state can prevent the unintended machinery failures. In rotating

N. N. Hung · J. Kim · J.-M. Kim (✉)
School of Computer Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

N. N. Hung
e-mail: hungnguyenvldt@gmail.com

J. Kim
e-mail: kjy7097@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_46

machinery, about 50% of faults are related to rolling element bearings [1]. These bearings are critical mechanical components which make the relative moment of systems smoother and support diametric and thrust loading [2]. Consequently, the bearing defects are the most frequent in among electrical and mechanical faults in machinery because they are very easily affected by operating environment condition, overloading, and misalignment [3]. Bearing defects are the primary cause of the suddenly mechanical breakdown in industry and lead to enormous economic losses. Thus, accurately evaluating the fault severity in bearings is urgently required for faults condition monitoring and diagnosis systems.

In this paper, we propose an efficient approach to evaluate the severe degree of the defects in bearing under variable crack size using the envelope power spectrum and the discrete wavelet packet transform (DWPT). Envelope power analysis is the most effective technique for signal demodulation, which detects the frequency periodic impulses showing the abnormal symptoms in the defective bearing [4]. However, these abnormal signatures of incipient bearing defects can occur in anywhere in the frequency domain of the envelope power spectrum. Therefore, the DWPT-based frequency decomposition is applied to analyze sub-band signal for extracting the intrinsic information about the bearing failures. Otherwise, there is still no general consensus on determining the most informative sub-band signal that is optimal for accurately fault evaluation at the early state in bearings. To solve this issue, kurtogram methods are extensively used in [5, 6] to select an optimal sub-band containing essential information about bearing faults. However, the drawback of this technique is not precisely proportional to the severe degree of defects in bearings because the sub-band signals can consist of the frequency components not caused by bearing failures such as the operating frequency and noise frequency components. This paper proposes an efficient evaluation methodology of the defect severity in bearing by using a Gaussian distribution model-based window for only capturing exactly characteristic defect frequencies in the envelope analysis domain [7]. Afterwards, a ratio of defect spectral component to residual spectral component (RDR) is estimated from the acquired envelope spectral signal that is an appropriate metric for evaluating the severity of defects in bearings. Through the RDR calculation for each sub-band in DWPT decomposition, the most optimal sub-band yielding the highest RDR value is selected for reliable assessment of bearing fault conditions.

The rest of this paper is presented as follows. Section 2 describes a data acquisition model and the bearing characteristic frequencies. Section 3 presents an effective methodology for evaluation of the fault severity in bearing, and Sect. 4 discusses the experimental results. Finally, Sect. 5 concludes this paper.

2 Data Acquisition

A machinery fault simulation is set up for evaluation of bearing failure severity, as shown in Fig. 1a. To capture the intrinsic information about faults, 5-s AE signals sampled at 250 kHz are obtained for each defect in bearing. In this paper, three

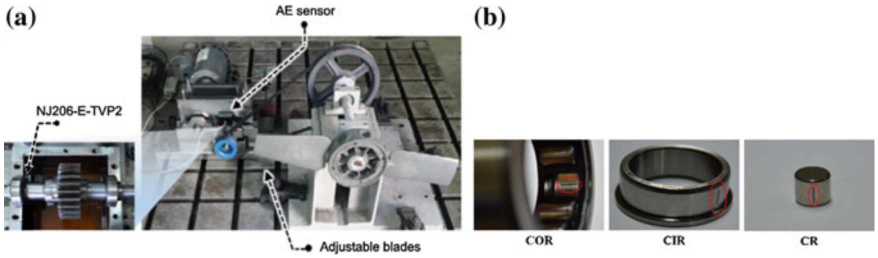


Fig. 1 **a** Data acquisition system and **b** primary bearing defect conditions

Table 1 Description of the characteristic defect frequency components

Crack size	Length (mm)	Width (mm)	Depth (mm)
	3	0.35	0.30
	12	0.49	0.50
Rotation speed	500 revolution per minute (rpm)		
Defect frequencies	BPFO = 43.68 Hz, BPFI = 64.65 Hz, 2xBSF = 41.44 Hz, and FTF = 3.36 Hz		

different primary fault types of a bearing are generated in the following locations: crack on outer-race (COR), crack on inner-race (CIR) and crack on a roller (CR), as illustrated in Fig. 1b. A detailed description of the bearing defect frequencies is given in Table 1, in which are ball pass frequency of the outer race (BPFO), ball pass frequency of the inner race (BPFI), and ball spin frequency (BSF), as defined in [8].

3 Evaluation Methodology of Bearing Fault Severity

The comprehensive evaluation methodology of the severity of defects for health condition monitoring in bearings is presented in Fig. 2. The input AE signal related to each bearing failure condition is first decomposed into a series of sub-bands by using the DWPT [9]. This paper employs 4-level decomposition with the Daubechies 4 (Db4) for analysis, result in a total of obtained $2^5 - 1$ sub-band signals. An effective envelope analysis is then applied to each sub-band signal to detect the characteristic defect frequencies in bearings.

In this paper, the sub-band signals decomposed by DWPT are processed further with envelope analysis to extract the frequency periodic impulses showing the abnormal symptoms of failures. The envelope power spectrum of each reconstructed sub-band signal $x_{sb}(t)$ is calculated by the Hilbert transform [10], as follows:

$$\hat{x}_{sb}(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x_{sb}(\tau)}{t - \tau} d\tau \tag{1}$$

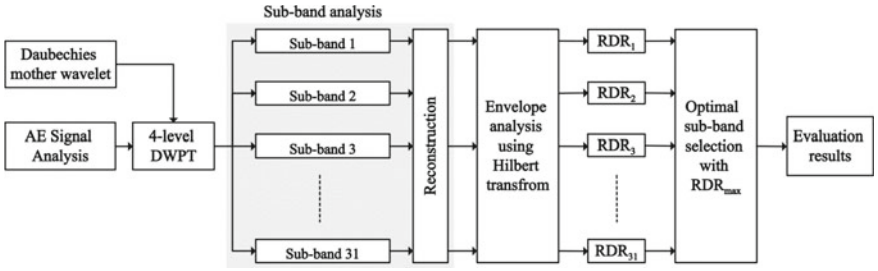


Fig. 2 An overall diagram for evaluation of the degree of defects in bearing

A combining the initial sub-band signal $x_{sb}(t)$ and its Hilbert transformed signal $\hat{x}_{sb}(t)$ is then performed to generate the analytical signal $s(t)$:

$$s(t) = x_{sb}(t) + i \cdot \hat{x}_{sb}(t),\, i = \sqrt{-1} \tag{2}$$

Based on the analytical signal, the envelope signal $x_{env}(t)$ is calculated:

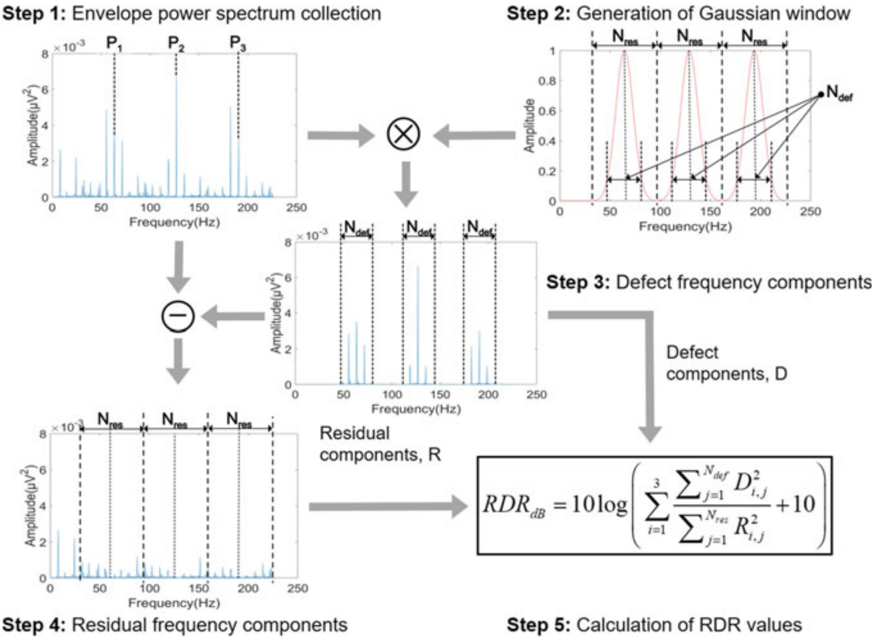
$$x_{env}(t) = |s(t)| = \sqrt{x_{sb}(t)^2 + \hat{x}_{sb}(t)^2} \tag{3}$$

Finally, the fast Fourier transform (FFT) for the envelope signal $F\{x_{env}(t)\}$ is realized and the envelope power spectrum $x_{eps}(t)$ is collected by squaring the absolute FFT values.

To correctly evaluate the severity of defects from the envelope power spectral signal of each sub-band, we calculate the RDR values in this paper. The RDR calculation is performed by generating a Gaussian window to obtain the defect frequency component of the BPFO, BPFI, BSF and their harmonics in the envelope power spectrum. The step by step of computation process is detailed in Fig. 3. The value of the window $w(j)$ is defined as follows:

$$w(j) = \begin{cases} \sum_{i=1}^{3} \exp\left(-\frac{1}{2}\left(\alpha \frac{(j-P_i)}{N_{res}/2}\right)^2\right), \\ 0, \quad otherwise \end{cases} \tag{4}$$

$P_i - f_{range} \leq j \leq P_i + f_{range}$,$P_i$ is the ith harmonic in spectral analysis of bearing defect frequency component ($i = 3$ in this paper), and N_{res} is the number of frequency bins around each harmonic of the defect frequencies, with: $N_{res} = 2 \cdot f_{range}/f_{resolution}$, where f_{range} defines the frequency range for calculation of the Gaussian window values, $f_{resolution}$ is the frequency resolution, and α is a parameter representing the distribution of Gaussian random variables. It is necessary to define an appropriate range of frequency for capturing intrinsic information about failures. Thus, a narrow frequency range with $f_{range} = 1/4\, BPFO$ is defined for assessing the outer failure, whereas a wide frequency range with $f_{range} = 1/2\, BPFI$ or BSF


Step 1: Envelope power spectrum collection
Step 2: Generation of Gaussian window
Step 3: Defect frequency components
Step 4: Residual frequency components
Step 5: Calculation of RDR values
P1
P2
P3
N_res
N_def
Defect components, D
Residual components, R
Amplitude(μV²)
Amplitude
Frequency(Hz)
×10⁻³
0
50
100
150
200
250
0.2
0.4
0.6
0.8
1
2
4
6
8

$$RDR_{dB}=10\log\left(\sum_{i=1}^{3}\frac{\sum_{j=1}^{N_{def}}D_{i,j}^{2}}{\sum_{j=1}^{N_{res}}R_{i,j}^{2}}+10\right)$$

Fig. 3 The overall process of the Gaussian window-based RDR calculation for each sub-band

is used for evaluating the inner and roller failures. Finally, the RDR value is calculated as follows:

$$RDR_{dB} = 10\log\left(\sum_{i=1}^{3}\frac{\sum_{j=1}^{N_{def}} D_{i,j}^{2}}{\sum_{j=1}^{N_{res}} R_{i,j}^{2}} + 10\right) \tag{5}$$

where N_{def} is the number of frequency bins containing the defect components around their harmonics. $D_{i,j}$ is the defect frequency components calculated by multiplying the window $w(j)$ with the envelope power spectrum $x_{eps}(t)$, and $R_{i,j}$ is the residual frequency component obtained by subtracting the defect frequency components.

After accurate estimation of the defective degree at different decomposition levels of DWPT by the RDR calculation for each obtained sub-band, an optimal sub-band containing the explicit signatures about failures, which yields the highest RDR value, is selected for reliable evaluation of the fault severity in bearings.

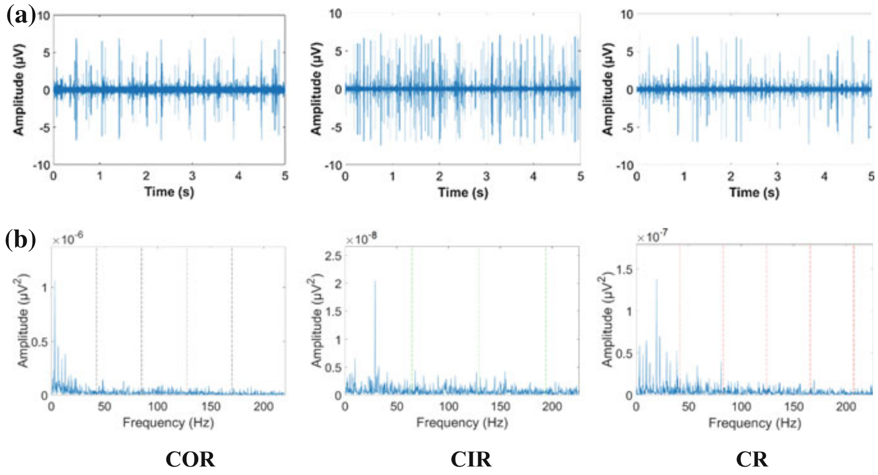
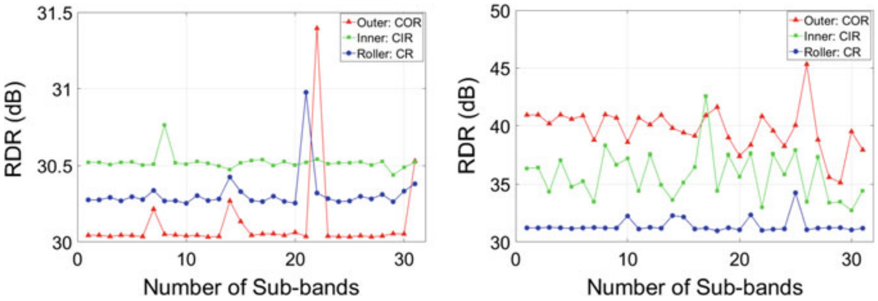


Fig. 4 **a** The acquired raw AE signals of various defect conditions at 3 mm crack size and **b** their original envelope power spectra

4 Experimental Results

This paper evaluated the fault severity in bearing via capturing the harmonics of the defect frequency components for each COR, CIR, and CR form envelope analysis. The 5-s AE signals of each defect are first decomposed to the sub-band signals using the DWPT. The bearing defects are detected by calculating the envelope power spectrum of $2^5 - 1$ sub-bands from 4-level DWPT with Db4 mother wavelet function. However, the abnormal symptoms of bearing defects can be revealed in anywhere in the sub-band signals. To exactly estimate the degree of defects in bearing from the obtained sub-bands, the RDR values are required for each sub-band by determining the appropriate Gaussian windows corresponding to the bearing fault conditions. The more clearly the spectral magnitudes around the harmonics of defect frequencies (BPFO, BPFI, and 2xBSF), the higher the RDR value. Figure 4 presents the acquired original AE signals of the different bearing fault conditions in case of 3 mm crack size and their initial envelope power spectra. Each fault condition represents a unique waveform in the time domain of outer, inner and roller. Based on envelope analysis of raw AE signals, it is difficult to correctly evaluate the defective degree in bearing because of unclearly presence of the characteristic defect frequencies from their original envelope power spectra.

In this study, the DWPT is successfully exploited for signal decomposition to obtain the sub-bands containing explicit abnormal signatures of bearing failures. To determine an optimal sub-band, we assess the degree of defects in a total of 31 the acquired sub-band signals by the RDR calculations under their envelope spectral signals, respectively. The most informatory sub-band corresponding to the highest RDR value is selected for correct assessment of the fault severity in bearing. The



(a) 3 mm **(b)** 12 mm

Fig. 5 Evaluating the defective degree of each sub-band at the different decomposition levels under variable crack sizes

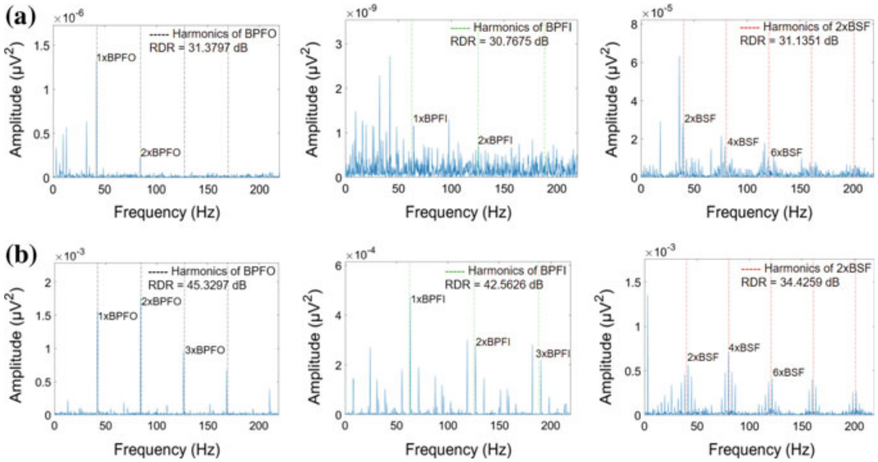


Fig. 6 Measurement of the bearing fault severity based on envelop power spectrum of optimal sub-bands in crack size of **a** 3 mm and **b** 12 mm

performance of the proposed methodology is demonstrated based on experimental results for evaluation of the primary defects caused by variable crack sizes (3 and 12 mm) on the surface of outer, inner, and roller in bearing. Figure 5 shows the degree of bearing defects (COR, CIR, and CR) for each sub-band at the different decomposition levels measured by the RDR calculation. The envelope power spectrum of optimal sub-band selected related to bearing conditions for precisely evaluating the fault severity is illustrated in Fig. 6, which show clearly the characteristic defect frequencies of bearing.

5 Conclusions

Accurately assessment of the severity of defects in bearings is immediately required for reliable health condition monitoring and fault diagnosis systems. This paper successfully exploits the DWPT and envelope power spectrum analysis in evaluating the severe degree of defect frequencies for COR, CIR, and CR in bearing under variable crack size. The envelope analysis is carried out for each sub-band signal to identify the characteristic defect frequency components of bearing. Furthermore, the RDR values are calculated effectively using appropriate Gaussian windows to determine an optimal sub-band showing explicitly intrinsic information about bearing failures. The RDR-based most informative sub-band selection yields significant effectiveness for precise evaluation of the severity of defect signatures in bearing.

Acknowledgements This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (Grant No. 20162220100050, No. 20161120100350, and No. 20172510102130), in part by The Leading Human Resource Training Program of Regional Neo Industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant No. NRF-2016H1D5A1910564), and in part by the Basic Research Program through the NRF funded by the Ministry of Education (Grant No. 2016R1D1A3B03931927).

References

1. Benbouzid MEH (2000) A review of induction motors signature analysis as a medium for faults detection. IEEE Trans Ind Electron 47:984–993
2. Leite VCMN, Silva JGBD, Torres GL, Veloso GFC, Silva LEBD, Bonaldi EL, Oliveira LEDLD (2017) Bearing fault detection in induction machine using squared envelope analysis of stator current. In: Darji PH (ed) Bearing technology. InTech, Rijeka, pp Ch 05
3. Kang M, Kim J, Kim JM (2015) An FPGA-based multicore system for real-time bearing fault diagnosis using ultrasampling rate AE SIGNALS. IEEE Trans Ind Electron 62:2319–2329
4. Leite VCMN, Silva JGBD, Veloso GFC, Silva LEBD, Lambert-Torres G, Bonaldi EL, Oliveira LEDLD (2015) Detection of localized bearing faults in induction machines by spectral kurtosis and envelope analysis of stator current. IEEE Trans Ind Electron 62:1855–1865
5. Wang D, Tse PW, Tsui KL (2013) An enhanced Kurtogram method for fault diagnosis of rolling element bearings. Mech Syst Signal Process 35:176–199
6. Zhang X, Kang J, Xiao L, Zhao J, Teng H (2015) A new improved kurtogram and its application to bearing fault diagnosis. Shock Vib 2015:22
7. Kang M, Kim J, Choi B-K, Kim J-M (2015) Envelope analysis with a genetic algorithm-based adaptive filter bank for bearing fault detection. J Acoust Soc Am 138, EL65–EL70
8. Bediaga I, Mendizabal X, Arnaiz A, Munoa J (2013) Ball bearing damage detection using traditional signal processing algorithms. IEEE Instrum Meas Mag 16:20–25
9. Yan R, Gao RX, Chen X (2014) Wavelets for fault diagnosis of rotary machines: a review with applications. Sig Process 96:1–15
10. Wang D, Miao Q, Fan X, Huang H-Z (2009) Rolling element bearing fault detection using an improved combination of Hilbert and wavelet transforms. J Mech Sci Technol 23:3292–3301

Event Extraction from Streaming System Logs

Shuting Guo, Zheng Liu, Wenyan Chen and Tao Li

Abstract Log data is typically the only available data source recording system health information. Event extraction converts unstructured log messages into structured event signatures. Existing methods, whether batch or streaming methods, require true event signatures to guide parameter selection. This paper presents a streaming event extraction method that eliminates the demands of external tags and generates appropriate event signatures by evaluating the quality of them. Experimental results show that our approach can parse log message into high-quality information efficiently and detect more anomalies.

Keywords Event extraction · Log parsing · Streaming system logs

1 Introduction

Big data mining tasks come from real applications and it can show its true value with specific application data and appropriate algorithms [1]. As one of the most popular tasks in data mining, log analysis is widely discussed in recent years. System logs reveal the behavior information of service operations, therefore analyzing and mining system logs are the key issues in service management to enhance system stability,

S. Guo · Z. Liu (✉) · W. Chen · T. Li
Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, People's Republic of China
e-mail: zliu@njupt.edu.cn

S. Guo
e-mail: 1016041128@njupt.edu.cn

W. Chen
e-mail: 1016041123@njupt.edu.cn

T. Li
e-mail: towerlee@njupt.edu.cn

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_47

security, and usability. Typical applications include anomaly detection [2, 3], fault diagnosis [4], association rule mining [5], and program verification [6].

Generally, system logs may contain millions of log messages, each of which is a text sequence corresponding to an output sentence in the source code. These system logs record important information about system operation statuses, such as job execution paths, resource usage patterns, program running exceptions, and other information [7]. Log messages could be divided into two parts: constant parts and variable parts. For the same kind of events, the constant parts of corresponding log messages have fixed expressions. While variable parts contain system running information such as running time, execution objects, and so on.

An event signature of a log is made up with constant parts and wildcard characters representing variable parts. Event extraction (i.e., log parsing) converts raw unstructured system logs into structured events with certain formats and then do some event mining tasks [8]. How to extract these events automatically, accurately and efficiently remains a challenging problem due to the following characteristics of system logs in practical applications:

- **Insufficient domain knowledge**. Traditional log parsing methods are based on handcrafted rules [3, 9]. In spite of the high parsing accuracy, it is usually very expensive to acquire sufficient domain knowledge because of the massive source codes from various software components with frequent system update [10].
- **Huge data volume**. In practical enterprise environments, the size of the accumulated logs could reach several terabytes per day, which is called "data lake" problem [11]. The huge data volume makes these unstructured logs unmanageable, and a significant amount of CPU time and memory spaces are required to process only a small part of them [10]. This is the main reason that most recent studies on offline log parsing [2, 12–14] are strict to the memory requirement of computers.

Recent studies on automatically event extraction have achieved quite good experimental results, e.g., LogSig [13] and Spell [15]. However, one major issue in these proposed solutions is that they need partial information from event signatures. LogSig requires an input parameter that is the exact number of event types, while in Spell, the parameters are carefully tuned based on the event characteristics such as classification threshold. However, in practical systems such information is usually not available and reachable.

The main contributions of this paper are summarized as follows:

- In this paper, we propose an automatic streaming event extraction method called BSG (basic signature generation) with only one parameter, which could be tuned by evaluating the event signature quality based on a sampling dataset. Compared with most existing log parsing approaches which require implicit event characteristics to guide parameter tuning, our approach can automatically parse logs into events in an unsupervised manner.
- We conduct extensive experiments on various log data sets from real-world domains, and carefully analyze the evaluation results, which demonstrate that our

proposed framework could achieve better performance than existing log parsing methods, regarding effectiveness and efficiency.

The reminder of this paper is organized as follows. In Sect. 2, we present related works about event extraction. In Sect. 3, we propose the detailed algorithm framework. We conduct extensive experiments and present the evaluation results in Sect. 4. Finally, we conclude the paper in Sects. 5 and 6.

2 Related Works

Earlier log parsing methods were developed for specific logs. Lang [9] uses regular expressions to parse logs, which need to be designed and maintained by the developer himself. Xu et al. [3] obtains the event signature of the log by analyzing the source code generated by the log. In recent years, researchers mainly consider automatic log parsing, which could be divided into either offline log analysis or streaming log analysis.

Offline log parsing. Offline log parsing method is a batch process that employs all the collected logs. Offline log parsing methods can be classified into three groups by motivation, based on heuristic rules like SLCT [12] and IPLoM [14], based on clustering such as LKE [2] and LogSig [13].

Online log parsing. Online log parsing operates in a streaming manner and does not require offline training step. Existing online log parsing methods are SHISO [16], Spell [15] and Drain [7]. SHISO mines and extracts log formats in real time with a structured tree based on nodes generated from log messages. Through comparing log with the currently generated event signature using the longest common subsequence, Spell can obtain the maximum matching sequence as the corresponding event signature. To speed up the matching process, Spell also uses the prefix tree to maintain the currently generated template for more efficient querying. Drain achieves event signatures using a fixed depth parse tree to parse system logs using manually designed rules.

3 Algorithm Overview

Suppose the system logs are coming in a streaming manner, our algorithm process could be divided into three steps: when a system log arrives, it is first preprocessed based on domain knowledge (details in Sect. 3.1), then it is grouped into bags of logs based on its length (details in Sect. 3.2) and finally it is compared with all current signatures to decide whether to merge it or not (details in Sect. 3.3).

3.1 Domain Knowledge Preprocessing

Domain knowledge related to system logs can help improve the log parsing accuracy [6]. Fields in these system logs recording attribute information might have a common expression. For example, “2017-12-01” and “2017-11-29” are both dates and “192.168.0.12” and “202.119.23.10” are both IP addresses. These fields might be shared in a wide range of system logs with various formats. The log generation mechanisms implicitly create associations between the terminologies and the situations. In the proposed framework, users can provide a list of regular expressions to map tokens to commonly used variable, which serves as the preprocessing rules. Note that the simple regular expressions used require much less human effort than those complex ones used in traditional methods to match the whole log messages. This is useful because developers tend to log certain properties in the same format in practice.

3.2 Bag of Logs Generation

Previous research efforts [7, 14] show that extracting events from logs of the same length could obtain a good analytical result. Similar to these research works, in this paper, we group log messages into bags of logs based on their length. All log messages are of the same length in the same bag. Note that there may have a fair amount of circumstances that one event signature could have logs of different lengths, e.g., “delete block blk [*]” and “delete block blk [*] blk [*]” for the same event. We will handle this case in later step by merging different bags.

3.3 Basic Signature Generation (BSG)

The event signature extraction process is essentially the longest common subsequence (LCS) problem, which has been proved as an NP-hard problem [13]. Therefore, it is impractical to find the longest common subsequence between log sequences. Because of the fixed length in each bag of logs defined before, we use edit distance instead of calculating longest common subsequence (LCS). The procedure of our proposed basic signature generation is shown in Fig. 1. During bags of logs generation, we create several objects called SigMaps, each of which has a unique log length. Similar to Spell [15], in each SigMap, we create a data structure called SigObj to hold currently parsed event signature and the related metadata information. Each SigObj contains a parsed template and a list of line indices that stores the line *ids* of the corresponding entries that lead to this signature. When a new log message arrives, we first calculate the length of its tokens and find the SigMap it belongs to. Then we generate the new signature by comparing this log message with all signatures in this

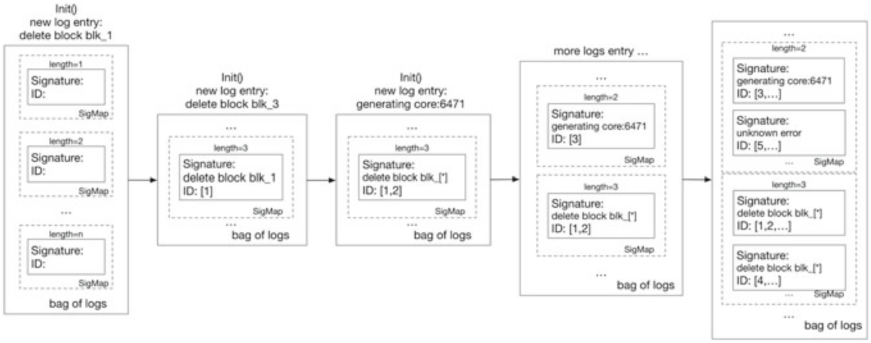


Fig. 1 Basic signature generation

SigMap. If the ratio of the number of constant parts in a new signature to the length of log message is greater than a threshold ε, we update the signature of corresponding SigObj and add the log message id. Otherwise, we create a new object, initialize its signature as this log message and insert it into the new object. Finally, we remove all the wildcards in logs and merge the event signature of different length of logs representing the same event.

Tuning Threshold with Explicit Characteristics Only

The tuning of this threshold ε is important to the parsing accuracy and the performance in applications. During basic event signature generation, the threshold ε leads the algorithm to build a new SigObj. The implicit event characteristics to which that most existing log parsing approaches tune their parameters according, such as real event types (ground truth) are not available and cannot guarantee the accuracy.

We employ similar strategy used in clustering validation [17] to tune the threshold in our procedures with explicit characteristics only. Similar to clustering validation, in basic signature generation, we want the logs with the same signature are more similar than logs with different signatures, that is, compactness and separation. Compactness measures how close related logs are and separation measures how well-separated are logs from different signatures.

Signature Compactness. The compactness of signature r is defined as $g(r)$, the ratio of c_r to l_r, where c_r denotes the number of constant tokens in event signature r, and l_r denotes the length of r. Compactness g measures the ratio of constant tokens of a signature, and the larger it is, the more compact the event signature is.

Signature Separation. Same like the separation within clusters, we use distance between logs of the same signature to measure the separation. The degree of separation between classes indicates the degree of looseness between classes, in intuition, the greater the distance between classes, the better the degree of looseness. Commonly used distances between texts include Jaccard distance and cosine distance. However, the different length of the signatures leads to difficulty in calculating distance. In addition, the same tokens in different locations may not necessarily show the same events, and the order of tokens has an effect on distance calculating. We use term pairs to solve the above two problems.

Term pairs can preserve the order information of message terms and have lower computation than the original log sequences. For example, there is a log template parsed from BGL:

$$\text{RAS KERNEL INFO generating core } [*]$$

We remove all parameters (marked as [*]) and extract each pairwise of terms and preserve the order of two terms. Then, the converted pairs are as follows:

{RAS, KERNEL}, {RAS, INFO}, {RAS, generating},
{RAS, core}, {KERNEL, INFO}, {KERNEL, generating},
{KERNEL, core}, {INFO, generating}, {INFO, core}, {generating, core}

We use Jaccard distance (as mentioned as $J(\cdot)$ in (1)) to measure distances among event signatures. Jaccard distance measures the degree of discrimination between two term pair sets using the ratio of the different elements of the two term pair sets to all the elements. We use the minimum Jaccard distance between term pairs to describe the degree of dispersion between the templates. In this log parsing task, we pay more attention to separation rather than compactness, we define clustering score *scluster* as below.

$$scluster(\varepsilon) = \sum_{i \in BSG(\varepsilon)} g(i) \sum_{i,j \in BSG(\varepsilon} J(tp(i), tp(j))^2 \quad (1)$$

We sample log data to select best threshold ε and get maximum value about *scluster*.

4 Experimental Evaluation

In this section, we present the experimental results on four real datasets to show both the effectiveness and the efficiency of our proposed log parsing method. All experiments are conducted on a Linux machine with a 24-core® Xeon® CPU E5-2420 v2 @ 2.20 GHz computer. All algorithms are implemented by Python and take the average results as the final results to avoid bias.

Table 1 provides a basic summarization of four datasets commonly used in previous research works [7, 13–15]. Those datasets collected from three kinds of systems cover a wide variety of structural formats and have complex structure.

We compare our algorithm with three methods IPLoM [14], LogSig [13] and Drain [7]. IPLoM is an offline method with 6 parameters to separate all logs into different clusters where all logs in the same cluster share same event signature. LogSig is

Table 1 Log datasets description

System	Number of logs	Tokens	Events	Description
BGL	4,747,963	10–102	376	BlueGene/L supercomputer
HDFS	11,175,629	8–29	29	Hadoop distributed file system
ThunderBird	96,080	1–124	66	Supercomputer
Apache error	99,997	10–20	40	Web server

another offline method which divides all logs into pre-defined number of clusters and searches the most representative signatures. Drain is a streaming method which utilizes a fixed depth parse tree to parse system logs using manually designed rules. For IPLoM, LogSig and Drain, we share the same parameter setting with [7] and [13]. BSG evaluates the quality of event signatures to find the best parameters for different datasets, which are 0.7 (BGL), 0.8 (HDFS), 0.3 (Apache Error) and 0.1 (Thunderbird).

4.1 Effectiveness of Different Methods

We evaluate the accuracy of different log parsing methods using F-measure on the sampled 2k data sets described in Table 1. From Table 2, we observe that our method outperforms other three algorithms in Apache Error and Thunderbird. LogSig performs relatively poorly on Apache Error and Thuderbird because it randomly allocates the logs to k groups based on the predefined number of events k, and then adjusts the group to which the log belongs by calculating the change in a potential function. The random initialization in LogSig brings certain impact to the result. IPLoM and Drain perform a little better than LogSig. Drain search by preceding tokens based on an assumption that in the beginning positions of a log message tokens are more likely to be constant parts. This leads to the poor performance of Drain on Apache Error dataset for which has some variable parts in the beginning. BSG needs no parameter tuning. This method generates event signature by distance matching according to the event signature quality assessment score based on a sampling dataset. But it could not solve the situation that different length logs correspond to same events, thus could not get the same ground truth as the datasets.

Table 2 Parsing accuracy of different log parsing methods

	Apache error	BGL	HDFS	Thunderbird
IPLoM	0.79	0.99	**1.0**	0.88
LogSig	0.74	0.99	0.99	0.85
Drain	0.79	0.99	**1.0**	0.88
BSG	**0.90**	0.99	**1.0**	**0.92**

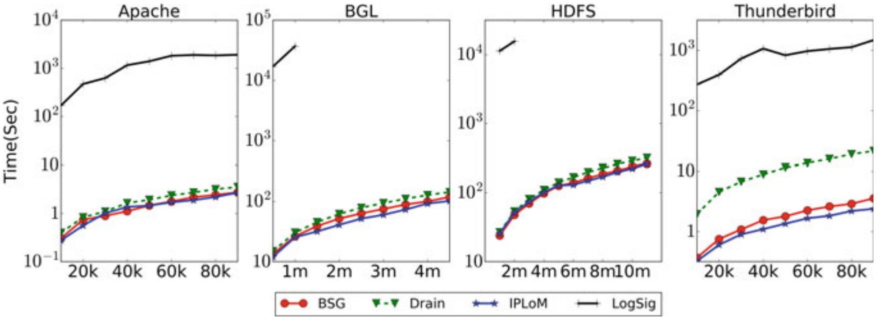


Fig. 2 Efficiency comparison of different methods

4.2 *Efficiency of Different Methods*

Figure 2 describes the efficiency of different methods. LogSig is the slowest algorithm for it's a clustering-based method with massive time consuming. Other three methods are equally matched because the time complexity of them is O(n). However, IPLoM requires a lot of I/O and intermediate storage, which makes it less scalable for big data in real systems. Besides, the adjustment of parameters is a time-consuming task. Drain needs to adjust the corresponding parameters for different data sets, and the parameters selection depends on the real ground truth. BSG can deal with those problem and shows good efficiency with O(n).

4.3 *Anomaly Detection Performance of Different Methods*

Like [7], we apply log parsing into anomaly detection with PCA tools [3]. Due to inefficiency and inaccuracy about LogSig, we compare IPLoM and Drain with BSG in HDFS and BGL datasets. In HDFS logs, there are 575,061 blockIDs, in which 16,838 are flagged as anomalies. BGL data contains 4,747,963 with 348,460 failures logs. We use one hour fixed windows and 10 min slide windows to slice logs as log sequences. In the anomaly detection task of HDFS, BSG reports 10,998 anomalies and detects 10,720 anomalies, which is the same as IPLoM and Drain. Table 3 shows

Table 3 Anomaly detection using different log parsing methods on BGL datasets

	Reported anomaly	Detected anomaly	False alarm
IPLoM	1,934	1,138 (55%)	796 (41%)
Drain	1,940	1,136 (55%)	804 (41%)
BSG	**2,125**	**1,249 (61%)**	**876 (41%)**

anomaly detection using different log parsing methods on BGL datasets, and we find that BSG gets the highest *Recall* among all methods.

5 Conclusion

In this work, we propose BSG, a streaming event extraction method that generates the appropriate event signatures by evaluating the quality of signatures to address the limitations of existing methods that need ground truth to guide parameter tuning. Experimental results show that our approach is simpler and has superior performance.

Acknowledgements This work is supported in part by Jiangsu Provincial Natural Science Foundation of China under Grant BK20171447, Jiangsu Provincial University Natural Science Research of China under Grant 17KJB520024, and Nanjing University of Posts and Telecommunications under Grant No. NY215045.

References

1. Li T, Liu Z, Zhou Q (2016) Application-driven big data mining. ZTE Technol J 22(2):49–52
2. Fu Q, Lou JG et al (2009) Execution anomaly detection in distributed systems through unstructured log analysis. In: 9th IEEE international conference on data mining. IEEE, pp 149–158
3. Xu W, Huang L et al (2009) Detecting large-scale system problems by mining console logs. In: 22nd ACM symposium on operating systems principles. ACM, pp 117–132
4. Nagaraj K, Killian C, Neville J (2012) Structured comparative analysis of systems logs to diagnose performance problems. In: 9th USENIX conference on networked systems design and implementation. USENIX Association, pp 26–26
5. Ma S, Hellerstein JL (2001) Mining partially periodic event patterns with unknown periods. In: 17th international conference on data engineering. IEEE, pp 205–214
6. Shang W, Jiang ZM et al (2013) Assisting developers of big data analytics applications when deploying on hadoop clouds. In: 35th international conference on software engineering. IEEE Press, pp 402–411
7. He P, Zhu J et al (2017) Drain: an online log parsing approach with fixed depth tree. In: 2017 IEEE international conference on web services. IEEE, pp 33–40
8. Liu Z, Li T, Wang J (2016) A survey on event mining for ICT network infrastructure management. ZTE Commun 14(2):47–55
9. Lang D (2013) Using sec. USENIX; Login Mag 38(6):38–43
10. Ning X, Jiang G, Chen H et al (2014) HLAer: a system for heterogeneous log analysis
11. Terrizzano IG, Schwarz PM et al Data wrangling: the challenging journey from the wild to the lake. In: 7th biennial conference on innovative data systems research

12. Vaarandi R (2003) A data clustering algorithm for mining patterns from event logs. In: 3th IEEE international workshop IP operations and management. IEEE, pp 119–126
13. Tang L, Li T, Perng CS (2011) Logsig: generating system events from raw textual logs. In: 20th ACM international conference on information and knowledge management. ACM, pp 785–794
14. Makanju A, Zincir-Heywood et al (2012) A lightweight algorithm for message type extraction in system application logs. IEEE Trans Knowl Data Eng 24(11):1921–1936
15. Du M, Li F (2016) Spell: streaming parsing of system event logs. In: 2016 IEEE 16th international conference on data mining. IEEE, pp 859–864
16. Mizutani M (2013) Incremental mining of system log format. In: 2013 IEEE International Conference on Services Computing. IEEE, pp 595–602
17. Liu Y, Li Z, Xiong H et al (2010) Understanding of internal clustering validation measures. In: 10th international conference on data mining. IEEE, pp 911–916

Anomaly Detection Using Agglomerative Hierarchical Clustering Algorithm

Fokrul Alom Mazarbhuiya, Mohammed Y. AlZahrani and Lilia Georgieva

Abstract Intrusion detection is becoming a hot topic of research for the information security people. There are mainly two classes of intrusion detection techniques namely anomaly detection techniques and signature recognition techniques. Anomaly detection techniques are gaining popularity among the researchers and new techniques and algorithms are developing every day. However, no techniques have been found to be absolutely perfect. Clustering is an important data mining techniques used to find patterns and data distribution in the datasets. It is primarily used to identify the dense and sparse regions in the datasets. The sparse regions were often considered as outliers. There are several clustering algorithms developed till today namely K-means, K-medoids, CLARA, CLARANS, DBSCAN, ROCK, BIRCH, CACTUS etc. Clustering techniques have been successfully used for the detection of anomaly in the datasets. The techniques were found to be useful in the design of a couple of anomaly based Intrusion Detection Systems (IDS). But most of the clustering techniques used for these purpose have taken partitioning approach. In this article, we propose a different clustering algorithm for the anomaly detection on network datasets. Our algorithm is an agglomerative hierarchical clustering algorithm which discovers outliers on the hybrid dataset with numeric and categorical attributes. For this purpose, we define a suitable similarity measure on both numeric and categorical attributes available on any network datasets.

Keywords Network data · Intrusion detection · Outlier analysis · Data instance
Multi-dimensional space · Cardinality of a set · Euclidean distance

F. A. Mazarbhuiya (✉) · M. Y. AlZahrani
Department of Information Technology, College of Computer Science & IT, Al Baha University,
Al Baha, Saudi Arabia
e-mail: fokrul_2005@yahoo.com

M. Y. AlZahrani
e-mail: imohduni@gmail.com

L. Georgieva
School of Mathematical & Computer Sciences, Heriot Watt University, Edinburgh, UK
e-mail: l.georgieva@hw.ac.uk

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_48

Canberra metric · Similarity of data instance pair · Similarity of clusters pair
Merge function

1 Introduction

Due to the widespread use of computer systems and associated network, it has become inexpensive to store, transfer and process the data. So there is a huge amount of data gathering every day. These data contains potentially useful information. The interpretation of such large amount of data and extracting the valuable knowledge from it is a challenging task. The word data mining is coined to describe the methods and techniques used for the task. Data mining is defined as the method of extracting non-trivial and previously unknown information or patterns in the data. There are several methods of data mining developed till today and clustering is one of them. Clustering is a data mining techniques based on unsupervised learning and is used to identify the data distribution and hidden patterns in the data. Clustering is mainly used to find out dense and sparse regions in the dataset. There are primarily two broad approaches of clustering namely partitioning approach and hierarchical approach. The hierarchical clustering approaches are mainly of two types (i) agglomerative clustering and (ii) divisive clustering. Lots of methods and algorithms have already been developed till today for clustering different types of data. In [1], authors have discussed the detail about clustering numerical data using distance function. In [2–4], authors, have presented methods for clustering categorical data. In [5], spatial data clustering is discussed.

An outlier is a data point which does not belong to any cluster. Finding outliers from large datasets is an interesting data mining problem and it has lots applications like fraud detections, anomaly detections, intrusion detections etc. In [6, 7], the methods for outliers detection are discussed. Intrusion is an attempt to compromise the integrity, confidentiality or availability of resources by accessing in an illegitimate way. There is a wide range of activities fall under this which includes denial of services, Probe, Remote to Local, User to Root. Detection and prevention of such activities is most hot topic of research now days. In [8, 9], the authors have discussed all the activities in details and proposed methods of intrusion detection using fuzzy clustering. Existing intrusion detection techniques may be broadly classified into—anomaly detection techniques and signature recognition techniques [10, 11]. In [12], authors have proposed a traffic anomaly detection method using K-means clustering algorithm. They have used weighted Euclidean distance in their proposed method. In [13], authors have proposed a statistical method based on dimensional reduction and pattern extraction for intrusion detection in wireless network. In [14], authors have proposed an anomaly detection method based on fuzzy c-means clustering. In their work they have considered the hybrid data consisting both numeric and categorical attributes and defined the distance formula in terms of distance formula on numeric attributes and dissimilarity formula on categorical attributes. Most of the work done so far has taken the partitioning clustering approaches and few were hierarchical.

In this paper, we propose an agglomerative hierarchical clustering algorithm for anomaly detection. First of all we define the similarity measure of two data instances as a weighted aggregate of similarity measure on their numeric attributes and that on their categorical attributes. The similarity measure on the numeric attributes is defined in terms of *Canberra metric* [15–18] and that on the categorical attributes is defined in terms of ratio of the cardinality of intersection of two data instances to that of the union of the same data instances. Next, we define the similarity of a pair of clusters say C_1 and C_2 (having $m1$ and $m2$ data instances respectively), as the weighted aggregate of the similarity measure on the numeric attributes and that on the categorical attributes of the data instances of C_1 and C_2. The similarity measure on the numeric attributes of C_1 and C_2 is given by *Canberra metric* [15–18] and is defined on the mean values of the numeric attributes for all data instances of C_1 and that of C_2. The similarity measure [19, 20] on the categorical attributes of C_1 and C_2 is defined as sum of the ratios of the cardinality of the pair-wise of intersections of the data instances of both C_1 and C_2 to the cardinality of their pair-wise union, then divide the factor by the product of the number data instances of C_1 and that of C_2 i.e. $(m1 \cdot m2)$ and then subtracting the whole factor from 1. Obviously, the value of the similarity measure (C_1, C_2), will be ranging from 0 to 1. For exactly similar data instances/clusters the value will be 0 and for exactly dissimilar its value will be 1. Then we define a *merge* function in terms of the similarity measure described above. Finally, an agglomerative hierarchical clustering algorithm for anomaly detection is presented in this paper. The algorithm is quite similar to one discussed in [21]. The algorithm supplies all the clusters along with a set of outliers. The extracted outliers are considered as anomalies.

The paper is organized as follows. In Sect. 2, we briefly discuss about the recent development in this field. In Sect. 3, we discuss the problem statement. The proposed algorithm for anomaly detection is discussed in Sect. 4. Finally, we conclude our paper with a brief conclusion given in Sect. 5.

2 Recent Works

Data mining has received a lot of attention to the researchers for its widespread uses. There are several methods of data mining available till today namely clustering, association rules mining, classification, sequential patterns mining. Out of these, clustering is one of the most popular techniques among the researchers. Clustering is an unsupervised data mining techniques and is mainly used for finding dense and sparse regions in the datasets. There are primarily two broad directions of clustering namely partitioning approach and hierarchical approach. In the partitioning approach data instances are divided into a predefined number clusters based on some criteria. In the hierarchical approach, the method sought to build a hierarchy of clusters one within another. The hierarchical clustering approaches are mainly of two type (i) agglomerative clustering techniques and (ii) divisive clustering techniques. So many methods and algorithms were developed till today for clustering different types of data. Clustering of numeric data is discussed in [1] where distance function is used as a criterion for clustering. In [2–4], authors, have presented methods for clustering categorical data. The spatial data clustering is discussed in [5]. A wavelet transform based method for clustering spatial data is discussed in [22]. Discovering outliers from large datasets is an interesting data mining problem that has been discussed in [6, 7].

Intrusion is an activity which can compromise the integrity, confidentiality or availability of resources by accessing in an illegitimate way. Intrusion activities and the methods of their detection is discussed in [8, 9]. In [12], authors have proposed a K-means algorithm based method for traffic anomaly detection using weighted Euclidean distance. In [13], authors have discussed a statistical approach based dimensional reduction and pattern extraction for intrusion detection in wireless network. A method based on K-means for anomaly detection in hybrid data is discussed in [14]. In [16], a nice algorithm for clustering categorical data is proposed which uses a similarity function as ratio of the cardinality of the intersection of attributes value to that of the union of the same. In [19, 20, 23], authors have used similar measure for clustering categorical data and documents data. In [23], the authors have defined the similarity measure on the clusters in terms fuzzy sets. In [18], authors have described *Canberra* metric based intrusion detection system. A Genetic algorithm based techniques for clustering mixed data is discussed in [24]. An intrusion detection algorithm based on the analysis of usage data coming from multiple partners is discussed [25] which reduces the number of false alarms. An intrusion detection method based on data mining techniques is also discussed in [26]. A method for network intrusion detection based on data mining techniques is discussed in [27, 28]. In [29], the authors have discussed an intrusion detection systems based on pattern recognition techniques. In [30], the authors have proposed an intrusion detection method based on fuzzy association rules and fuzzy frequency episodes. In [31], the authors have conducted a comparative study on different anomaly detection schemes in network intrusion detections. In [32], authors have proposed a method which facilitates the investigation of huge amounts of intrusion detection alerts by a specialist. Their approach has made use of process mining techniques to find attack strategies observed in intrusion alerts. In [33], the authors have extended the work of [32] by proposing an alert correlation approach with emphasis on visual models to assist network administrators in the investigation of multistage attack strategies. In [34], the authors have discussed a method of reduction of intrusion detection alarm based on root cause analysis and clustering. In [35], a method of finding alert correlation based on frequent itemset mining is discussed. In [36], the authors have proposed a method of mining for causal knowledge automatically based on the Markov property for identifying multi-stage attack. In [37], the authors have discussed an intrusion detection method using Support Vector Machine which reduces the time to build classification model and increases accuracy. An intrusion detection system based on classification is discussed in [38]. An intrusion detection method based on artificial neural network is discussed in [39, 40]. In [41], the authors have proposed a real-time framework for identifying multi-stage attack scenarios from alerts generated by IDS which uses sequential pattern mining. In [42], the authors have proposed an SVM-based intrusion detection system, which combines a hierarchical clustering algorithm and SVM. They have used the dataset KDD Cup 1999 to check the efficacy of their system by comparing its performances with others intrusion detection systems. In [43], the authors have introduced a regression based online method for anomaly detection of smart grid data. In [44], authors introduced a method based on the analysis of mutual dependencies of the separate slices of network which is used to detect unpredicted activity of user or network equipment. An Anomaly based Intrusion Detection System that can detect various network attacks is discussed in [45].

3 Problem Statement

Here our aim is to cluster the data having both numeric and categorical attributes. For example, each connection instance of KDD Cup 1999 network data has 41 properties with 3 flag properties and 38 numeric properties. However, most of the current works were directed towards numeric properties and a very little works were done in dealing with categorical properties.

If we consider both numeric and categorical attributes of the sample data, then we have to treat it as a hybrid data or mixed data. The similarity measure must be defined in terms of both the attributes. As our approach is agglomerative hierarchical, the distance function [14, 24], defined in terms Euclidean distance and dissimilarity cannot be used. Instead we take another measure call *similarity measure*. In below we describe some of the definitions used in the paper.

Definition 1 *(Similarity between a pair of connection instances)*

Let A and B are two data instances with dimension n, where first k-attributes are numeric and rest of the $(n-k)$ are categorical. The k-numeric attributes of A, B are denoted by A^{n}, B^{n} respectively, and $(n-k)$ categorical attributes are denoted by A^{c}, B^{c} respectively. The similarity measure between A and B, denoted by $S(A, B)$ is given by the expression.

$$S(A, B) = \frac{k \cdot S_1 + (n-k)\, S_2}{n} \tag{1}$$

Where $S_1 = S(A^{\mathrm{n}}, B^{\mathrm{n}})$ = similarity of A and B defined on the k-numeric attributes. Now, to find the expression for the similarity between two numeric variables, we proceed as follows.

Let $x = (x_1, x_2,\ldots, x_{\mathrm{k}})$ and $\mathrm{y} = (y_1, y_2,\ldots, y_{\mathrm{k}})$ two k-dimensional vectors. Then the *Canberra metric* [15–18], $d(x, y)$ is given by the formula.

$$d(x, y) = \sum_{i=1}^{k} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{2}$$

The range of the above *Canberra metric* [15–18] is [0, k]. To make it, [0, 1] we divide the Eq. (2) by k. Therefore, we get new formula for *Canberra metric*, which is our similarity measure S_1 and is expressed as below.

$$S_1 = \frac{1}{k}\sum_{i=1}^{k} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{3}$$

Again $S_2 = S(A^c, B^c)$ = similarity of A and B defined on rest of the $(n - k)$ attributes. For S_2, we use a formula quite similar to the formula given in [19, 20, 23]. The formula is given as below.

$$S_2 = 1 - \tfrac{|A^c \cap B^c|}{|A^c \cup B^c|} \tag{4}$$

Obviously the value of S_2 will be the ranging from 0 to 1. For, exactly similar categorical values of both the data instances, $A^c = B^c$, $S_2 = 0$, and for exactly, different data instances, $A^c \bigcap B^c = \phi$, so that $A^c \bigcap B^c = 0$, and thus $S_2 = 1$.

With help of the Eqs. (3) and (4), the Eq. (1) will give us the similarity value for the two data instances A and B. Here we write the formula as a weighted aggregate of the both numeric and categorical variables so that both the attributes will have proportional contribution on the similarity function and its value will be ranging from 0 to 1. Using the Eq. (1), we say that the two data instances A and B are similar if and only if $S(A, B)$ is less than or equal to a pre-assigned threshold value, otherwise they are dissimilar. Obviously they will be precisely similar for the value 0 and precisely dissimilar for the value 1.

Definition 2 *(Similarity between a pair of clusters)*
Let $C_1 = \{A[i]; i = 1, 2, \ldots, m_1\}$ and $C_2 = \{B[j]; \mathrm{i} = 1, 2, \ldots, m_2\}$ be two clusters having m_1 and m_2 data instances respectively, the similarity function $S(C_1, C_2)$ is given by the formula

$$S(C_1, C_2) = \frac{k \cdot S_1(\overline{A^n}[i], \overline{B^n}[j]) + (n - k) \cdot S_2(A^c[i], B^c[j])}{n} \tag{5}$$

where $S_1(\overline{A^n}[i], \overline{B^n}[j])$ is the *Canberra metric* defined on the arithmetic means of the numeric variables of the data instances $A[i]$ and $B[j]$ of C_1 and C_2 respectively and is expressed as

$$S_1(\overline{A^n}[i], \overline{B^n}[j]) = \frac{1}{k}\sum_{i=1}^{k} \frac{|\overline{x}_i - \overline{y}_i|}{|\overline{x}_i| + |\overline{y}_i|} \tag{6}$$

where $\overline{x}_i$ = arithmetic mean of the variable x of $A[i]$, $i = 1, 2, \ldots, m_1$, and $\overline{y}_i$ = arithmetic mean of the variable y of $B[i]$, $i = 1, 2, \ldots, m_2$.

Similarly, $S_2(A^c[i], B^c[j])$ is the similarity defined on the categorical attributes of the above-mentioned data instances of C_1 and C_2 and $S_2(A^c[i], B^c[j])$ is expressed by the formula

$$S_2(A^c[i], B^c[j]) = 1 - \frac{1}{m_1 \cdot m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{|A^c[i] \cap B^c[j]|}{|A^c[i] \cup B^c[j]|} \tag{7}$$

Here $A^c[i] \cap B^c[j]$ =pair-wise intersections of data instances of C_1 and that of C_2. Obviously there will be $m_1 \cdot m_2$ number of the ratios $\frac{|A^c[i] \cap B^c[j]|}{|A^c[i] \cup B^c[j]|}$ and each having maximum value 1. The lowest value of $S_2(A^c[i], B^c[j]) = 1 - \frac{m_1 m_2}{m_1 m_2} = 0$. Similarly, the highest value of $S_2(A^c[i], B^c[j]) = 1$. Thus the values of $S_2(A^c[i], B^c[j])$ ranges from 0 to 1.

With the help of Eqs. (6) and (7), Eq. (5), gives the similarity of clusters or inter-clusters similarity $S(C_1, C_2)$. Obviously its value ranges from 0 to 1. For exactly similar cluster-pair its value is 0 and for exactly different cluster-pair its value is 1.

Definition 3 *(Merger of Cluster)*

Let C_1 and C_2 be the two clusters having m_1 and m_2 data instances respectively. Let C be the cluster formed by merging C_1 and C_2. Then the *merge*() function is defined as $merge(C_1, C_2) = C1 \bigcup C2$, if and only if $S(C_1, C_2) \leq \sigma$, a pre-defined threshold.

4 Proposed Algorithm

At the beginning of the clustering process, each data instance is allocated to a separate cluster. Thereafter for every pair of data instances the *similarity measure* is computed and then the *merge* function is used to obtain larger clusters if and only the *similarity* value is found to be within certain limit (the definition of *similarity measure* and *merge* function is given in Sect. 3). At any level, for any two clusters say C_1 and C_2 (having m_1 and m_2 data instances respectively), the *similarity* value is calculated using the formula given in Sect. 3, to check whether they can be *merged* or not. If the *similarity* value is found to be within a certain pre-determined threshold then C_1 and C_2 are *merged* using *merge* function to form a new bigger cluster. The process of *merging* continuous till no *merger* is possible or there is only one cluster at the top. In bellow we present the pseudo code for the proposed algorithm.

```
Algorithm DataInstanceClustering(n, σ)

Input:   The number of data instances $A$[i]; i = 1, 2,…$n$, and threshold σ
Output:  A set of cluster $S$
Step1.   The set of clusters $S$, where each cluster $C$ of $S$ having one data instance
         $A[i]$
Step2.   If for any cluster $C_1 \in S$ and $S(C_1, C) \leq \sigma$, then $merge(C_1, C)$ to form a
         new cluster $C_2$ consisting of $C_1$ and $C$.
```

```
Step3.  Remove $C_1$ from $S$.
Step4.  Continue Step2 and Step3 till no merger of clusters is possible.
Step5.  Return $S$
Step6.  Find all outliers from $S$.
Step7.  Stop
```

The algorithm supplies the set of clusters S of data instances which also includes outliers that means the data instances which do not belong to any larger clusters. The outliers can be used to find anomalies among the data instances. The patterns obtained by the above algorithm can be used in designing an efficient Intrusion Detection System (IDS).

5 Conclusion

The anomaly detection technique is one of the techniques used for developing Intrusion Detection Systems. In this paper, we propose an agglomerative hierarchical clustering algorithm for anomaly detection. For clustering purpose, we define a suitable *similarity measure* in terms of similarities in both numeric and categorical attributes. The *similarity* on the numeric attributed is defined in terms *Canberra metric* and that on the categorical attributes is defined in terms of the ratio of the cardinality of intersection of two data instances to that of the union of the same data instances Then, we take the weighted average of the both value to find *similarity* of the data instances. Then data instances are merged based the similarity value to find larger clusters. At any level, a pair of clusters is merged based its *similarity* value defined in Sect. 3. The process continues till no cluster merging is possible. The algorithm gives as output, a set of clusters of data instances each cluster having similar type of instances. Obviously any data instance deviating from pattern extracted by the method would be an anomaly which is used in the design of an efficient Intrusion Detection System.

References

1. Hartigan JA (1975) Clustering algorithms. Wiley
2. Gibson D, Kleinberg J, Raghavan P (1998) Clustering categorical data: an approach based on dynamical systems. In: Proceedings of the 24th international conference on very large databases, New York, pp 311–323
3. Ng RT, Han J (1994) Efficient and effective clustering methods for spatial data mining. Santiago, Chile, In Proc. of the VLDB Conf, pp 144–155
4. Ganti V, Gehrke J, Ramakrishnan R (1999) CACTUS-clustering categorical data using summaries. In: Proceedings of the international conference on knowledge discovery and data mining, San Diego, CA, USA, pp 73–83
5. Guha S, Rastogi R, Shim K, Rock (1999) A robust clustering algorithm for categorical attributes. In: Proceedings of the IEEE international conference on data engineering, Sydney, pp 512–521

6. Pamula R, Deka JK, Nandi S (2011) An outlier detection method based on clustering. In: Proceedings of 2011 second international conference on emerging applications of information technology, India, Feb 2011, pp 253–256
7. Zhang Y, Liu J, Li H (2010) An outlier detection algorithm based on clustering analysis. In: The proceedings of 2010 first international conference on pervasive computing, signal processing and applications, China, Sept 2010
8. Sharma D (2011) Fuzzy clustering as an intrusion detection technique. Int J Comput Sci Commun Netw 1(1), 69–75
9. Xie L, Wang Y, Chen L, Yue G (2010) An anomaly detection method based on fuzzy c-means clustering algorithms. In: Proceedings of the second symposium on networking and network security, China, pp 89–92
10. Debar H, Dacier M, Wespi A (1999) Towards a taxonomy of intrusion detection systems. Comput Netw 31:805–822
11. Escamilla T (1998) Intrusion detection: network security beyond the firewall. Wiley, New York
12. Munz G, Li S, Carle G (2007) Traffic anomaly detection using k-means clustering. Allen Institute for Artificial Intelligence
13. Haldar NA, Abulaish M, Pasha SA (2012) A statistical pattern mining approach for identifying wireless network intruders. In: Advances in Intelligent Systems and Computing: Preface, July 2012, pp 131–140
14. Linquan X, Ying W, Liping C, Guangxue Y (2010) An anomaly detection method based on fuzzy c-means clustering algorithm. In: Proceedings of the second international symposium on networking and network security, China, Apr 2010, pp 089–092
15. Lance GN, Williams WT (1966) Computer programs for hierarchical polythetic classification ("similarity analysis"). Comput J 9(1):60–64
16. Lance GN, Williams WT (1967) Mixed-data classificatory programs I agglomerative systems. Aust Comput J 15–20
17. Clifford TH, Stephenson W (1975) An introduction to numerical classification. Academic Press. New York, San Fransisco, London
18. Emran SM, Ye N (2001) Robustness of Canberra metric in computer intrusion detection. In: Proceedings of 2001 IEEE workshop on information assurance and security, US Military Academy, NY, June 2001, pp 80–84
19. Dutta M, Mahanta AK, Mazumder M (2001) An algorithm for clustering of categorical data using concept of neighours. In: Proceedings of the 1st national workshop on soft data mining and intelligent systems, Tezpur University, India, pp 103–105
20. Dutta M, Mahanta AK (2006) An algorithm for clustering large categorical databases using a fuzzy set based approach. In: Proceedings national workshop on trends in advanced computing, Tezpur University, India
21. Mazarbhuiya FA, AlZahrani MY (2017) An efficient method for clustering periodic patterns. In: Computing conference 2017, SAI Conference, London, UK
22. Sheikholeslami G, Chatterjee S, Zhang A (1998) WaveCluster: a multi-resolution clustering approach for large spatial databases. In: Proceedings of 24th VLDB conference, New York, USA
23. Thaoroijam K, Mahanta AK (2016) A fuzzy based document clustering algorithm. Int J Comput Appl (0975–8887) 151(10):21–24
24. Li J, Gao XB, Jiao LC (2004) A GA-based clustering algorithm for large datasets with mixed numerical and categorical values. J Electron Inf Technol 26(8):1203–1209
25. Bama SS, Ahmed MSI, Saravanan A (2011) Network intrusion detection using clustering: a data mining approach. Int J Comput Appl 30(4):14–17
26. Lee W, Stolfo SJ (1998) Data mining approaches for intrusion detection. In: 7th conference on USENIX security symposium
27. Dokas P, Ertos L, Kumar V, Lazarevic A, Srivastava J, Tan PN (2002) Data mining for network intrusion detection. In: Proceedings of the NSF workshop on next generation data mining, Nov 2002

28. Bloedorn E, Christiansen AD, Hill W, Skorupka C, Talbot LM (2001) Data mining for network intrusion detection: how to get started. Technical report, MITRE
29. Esposito M, Mazzariello C, Oliviero F, Romano SP, Sansone C (2005) Evaluating pattern recognition techniques in intrusion detection systems. In: Proceedings of the 5th international workshop on pattern recognition in information systems (PRIS) 2005, May 2005, pp 144–153
30. Luo J, Bridges S (2000) Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. Int J Intell Syst 15(8):687–704
31. Lazarevic A, Ertöz L, Kumar V, Ozgur A, Srivastava J (2003) A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the third SIAM international conference on data mining, May 2003
32. Alvarenga SC, ZarpelÃƒÂ£o BB, Junior SB, Miani RS, Cukier M (2015) Discovering attack strategies using process mining. In: The eleventh advanced international conference on telecommunications, AICT 2015, IARIA, pp 119–125
33. de Alvarengaa SC, Juniora SB, Mianib RS, Cukierc M, Zarpelãoa BB (2017) Process mining and hierarchical clustering to help intrusion alert visualization. Comput Secur
34. Al-Mamory SO, Zhang H (2009) Intrusion detection alarms reduction using root cause analysis and clustering. Comput Commun 32(2):419–430
35. Lagzian S, Amiri F, Enayati A, Gharaee H (2012) Frequent item set mining-based alert correlation for extracting multi-stage attack scenarios. In: 2012 sixth international symposium on telecommunications (IST). IEEE, pp 1010–1014
36. Xuewei F, Dongxia W, Minhuan H, Xiaoxia S (2014) An approach of discovering causal knowledge for alert correlating based on data mining. In: 2014 IEEE 12th international conference on dependable, autonomic and secure computing (DASC). IEEE, pp 57–62
37. Bhavsar YB, Waghmare KC (2013) Intrusion detection system using data mining technique: support vector machine. Int J Emerg Technol Adv Eng 3(3):581–586
38. Wankhede R, Chole V (2016) Intrusion detection system using classification technique. Int J Comput Appl (0975–8887) 139(11):25–28
39. Shun J, Malki HA (2008) Network intrusion detection systems using neural network. In: ICNC 2008. IEEE Explore
40. Poojitha G, Kumar KN, Reddy RJ (2010) Intrusion detection using artificial neural network. In: Proceedings of ICCCN 2010. IEEE Explore
41. Bahareth FA, Bamasak OO Constructing attack scenario using sequential pattern mining with correlated candidate sequences. Res Bull Jordan ACM, II(III):102–108
42. Horng SJ, Su MY, Chen YH, Kao TW, Chen RJ, Lai JL, Perkasa CD (2011) A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst Appl 38(1):306–313
43. Liu X, Nielsen PS (2016) Regression-based online anomaly detection for smart grid data. Technical University of Denmark, Kgs. Lyngby, Denmark
44. Gladkykh T, Hnot T, Solskyy V (2016) Fuzzy logic inference for unsupervised anomaly detection. In: IEEE first international conference on data stream mining & processing, 23–27, pp 42–47
45. Mane VD, Pawar SN (2016) Anomaly based IDS using back propagation neural network. Int J Comput Appl (0975–8887) 136(10):29–34

User Identification on Social Networks Through Text Mining Techniques: A Systematic Literature Review

Kinza Zahra, Farooque Azam, Wasi Haider Butt and Fauqia Ilyas

Abstract Social connection between the set of people is known as social network analysis. People keep numerous identities on various online social sites. User-related network data has distinctive information which shows user interests, behavioral patterns, and political views. By using these behaviors individually and collectively are of great help to recognize users across social networks. SLR (Systematic Literature Review) has been performed to distinguish 31 papers published during 2010–2018. The idea is to determine user identification categories that are used to classify users. Furthermore, to identify algorithms, models, methods, and tools that has been suggested since 2010 for user characterization. We have identified 10 algorithms, 19 models, 5 methods and 8 tools that have proposed for 5 user identification categories. Finally, we empirically evaluated that text mining techniques are promising approaches for the identification of users on online social networks.

Keywords User identification · Online social networks · Text classification
Text mining · Tools

K. Zahra (✉) · F. Azam · W. H. Butt · F. Ilyas
Department of Computer Engineering, College of E&ME,
National University of Sciences and Technology (NUST), Islamabad 12, Pakistan
e-mail: kinza.zahra15@ce.ceme.edu.pk

F. Azam
e-mail: farooq@ceme.nust.edu.pk

W. H. Butt
e-mail: wasi@ceme.nust.edu.pk

F. Ilyas
e-mail: fauqia.ilyas85@ce.ceme.edu.pk

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_49

1 Introduction

OSN (Online Social Networks) such as Facebook, Twitter, and Reddit, etc. have become extremely popular over the past decade and been one of the most common communication tools [1]. To integrate these OSN sites, for social networks it is essential to discover the identity of a user [2]. User identification based on text has attracted the attention of many types of research. User identification from text related to behavioral patterns [3], demographic characteristics of authors like age and gender [4] is essential in forensics, security, and advertisement. For instance, one would like to learn about the behavior of the author of aggressive and criminal textual message, or organizations might be intrigued to find out about demographic attributes of individuals who like or dislike their items, given the web journals and online product analysis.

Prevailing literature can be categorized into two groups, i.e., Systematic and Traditional literature reviews. State-of-the-art work and current research trends are mainly covered by traditional literature reviews while the focus of systematic literature reviews is to provide solutions to the research questions involving user identification. Despite a large number of empirical studies on user identification genres, models and algorithms there was a need to combine all these frameworks of same domain in one literature so that they can be compared regarding performance, accuracy, and validation against state-of-the-art machine learning algorithms. As far as we know, No systematic literature review can be found that concentrates on online user identification through text mining techniques, which encourages and stimulate our efforts in this study. The purpose of this SLR is to provide solutions to the following research questions:

RQ1: Which of the user identification categories are mainly focused on researching online social networks through text classification?

RQ2: What are the models and algorithms that are used to identify user through text classification?

RQ3: What are the current tools for user identification in text classification research area?

The paper is structured as follows. Section 2 illustrates the methodology performed in this paper. Section 3 demonstrates and shows the results. Section 4 presents the discussion and limitations. Conclusion and future work are suggested in Sect. 5.

2 Methodology

This paper was commenced as a systematic literature review established on the primitive directions as presented by Kitchenham [5]. It is intended to improve, assess, and understand the accessible material regarding research contributions that must be examined applicable concerning user identification and commensurate with already stated research questions. To lessen at least the possibility of results being writer's partial preferences, a systematic literature review is carried out on holding to extensive and promptly established guidelines.

2.1 Review Protocol Development

Our review protocol is accomplished under already defined guideline by Kitchenham [5]. In first stage, five digital libraries IEEE Explorer, ACM Digital Library, Springer, Elsevier and Taylor & Francis were selected to identify relevant papers. Our review protocol defines the criteria of inclusion and exclusion, search strategy, evaluation of quality and extraction and synthesis of data.

2.1.1 Inclusion and Exclusion Criteria

Our inclusion criteria concentrated on high-level comprehensive papers addressing original work on user characterization. We established the following selection requirements for inclusion criteria:

- Using text mining techniques for user identification.
- Using Machine Learning techniques for preprocessing and feature selection.
- Described categories, i.e., user behavior, attribute identification and spammers for user identification.
- Study with both journal and conference type, solely journal type is included.

Papers having these features were excluded:

- Papers that are related to user characterization other than text.
- Duplicated studies (only one copy of each study was included).

2.1.2 Search Process

Search process of our literature review commenced with the selection of digital libraries and research questions. This selection plays a key part to show the comprehensiveness and completeness of accumulated papers as illustrated in Table 1. Following steps are accomplished in the search process as shown in Fig. 1.

- Digital libraries (IEEE, SPRINGER, ELSEVIER, ACM, Taylor, and Francis) are focused.
- Journals and conference papers strain by title, keywords, and abstract by using the criteria of inclusion and exclusion.
- Considerable search terms were acquired from RQ’s.
- Boolean AND was applied to restrict the search.

Table 1 Selection criteria

Sr. #	Search terms	Operator	IEEE	Springer	ACM	Elsevier	Taylor and Francis
1	Online user characterization	AND	154	1915	14105	3043	64
2	Online user categorization	AND	131	1899	95	2840	1556
3	Social network analysis	AND	12853	32893	4225	22694	1348
4	Online social networks	AND	14780	15792	265	15106	2264
5	Models for OSN's	AND	5560	13	102	13463	1971

Table 2 Quality assessment checklist

Sr. #	Questions	Answer
1	Does the paper aim to identify user through text categorization?	Yes
2	Does the research review any of the preceding papers?	Yes
3	Do the findings address the original research questions?	Yes
4	Is the paper biased towards one user identification algorithm, model or approach?	No

2.1.3 Quality Assessment

We evaluated the quality assessment criteria described in the study performed by Kitchenham [5]. Quality assessment checklist was amplified to evaluate the quality of the selected papers as shown in Table 2.

2.1.4 Data Extraction and Synthesis

We identified Data Extraction and Synthesis to outline extraction structures to precisely record and gather the data by studying chosen publications. We formulated excel sheet to precisely record data to provide the solution to the questions. We included general data about the paper, for example, title, the name of the author, publication year, research type, and overall summary. While in data synthesis each question was separately evaluated against the results. Results in tabular form are illustrated in Table 3.

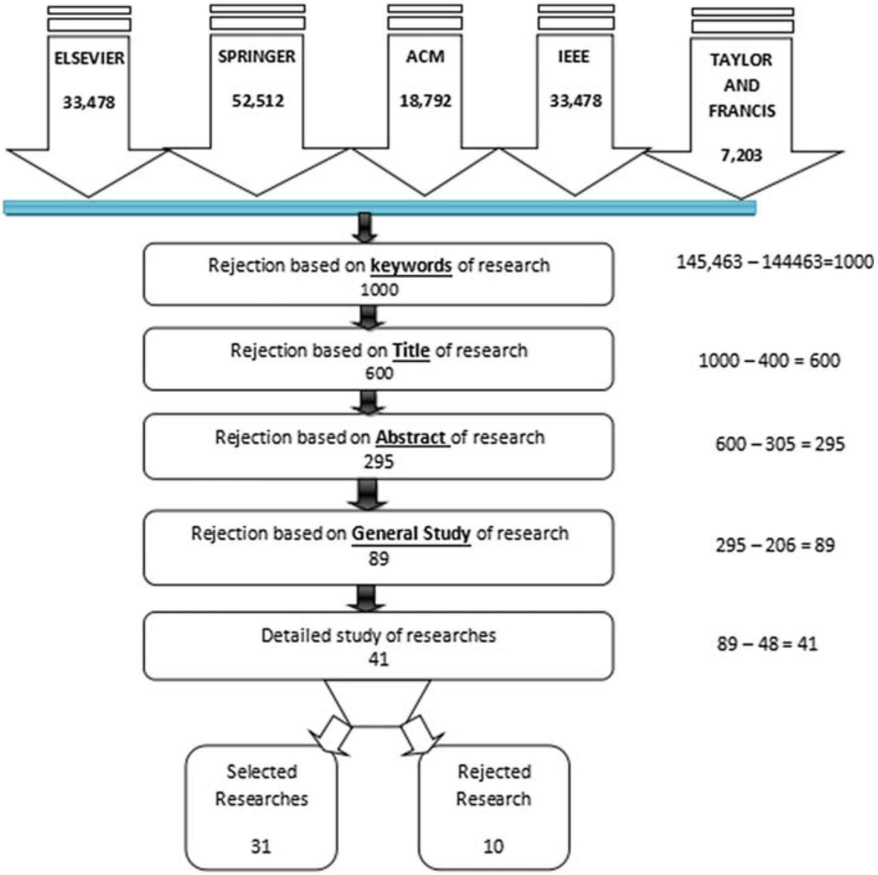


Fig. 1 Search process

3 Results

Findings of the systematic literature review by determined review protocol are explored in this section. Figure 2 shows a graphical depiction of the selected papers from the years 2010 to 2018 contemplating all researches by year. Results are then analyzed by quantitative and cross-examination techniques.

3.1 General View of Selected Studies

We identified 31 studies in the field of machine learning based user categorization. The papers were published between years 2010 to 2018. From selected studies,

Table 3 Data extraction and synthesis

Sr. #	Description	Details
1	Bibliographic information	Author, title, publication year, publisher details, and type of research(i.e. journal or conference)
Extraction of data		
2	Overview	Main objective of the selected paper
3	Results	Results acquired from the selected paper
4	Data collection	Qualitative and quantitative method used
5	Assumptions	To validate the outcome
6	Validation	State-of-the-art ML models are used to validate results
Synthesis of data		
7	Assigning user categories	Categories for user identification based on text classification
8	Framework selection	Models, algorithms and methods for user identification based on text classification
9	Tool selection	Emphasize on tools used for user identification based on text classification

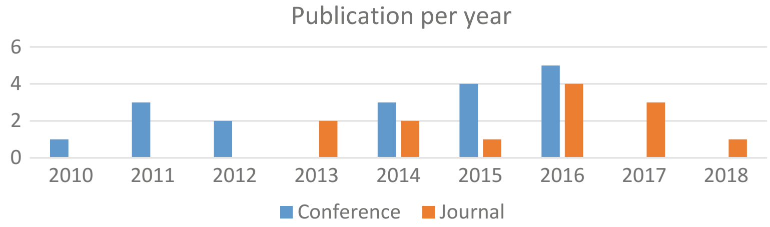


Fig. 2 Number of selected researches per year

13 papers were published in the journal, while 18 papers appeared in conference proceedings. Regarding the types of studies, all the selected studies are from experimental research. Table 4 shows overview of the selected work.

3.2 User Identification Categories (RQ 1)

All papers have been analyzed to find out from which user identification category authors are contributing research results on text in particular. User identification categories, i.e., behavioral, attribute, topic, spam, and crime are discussed in RQ1. Among all the above categories behavioral and attribute are two most frequently used categories, they together were mentioned by 65% of the selected papers as illustrated in Table 5. Compared to other identification categories, behavioral and attribute seem

Table 4 Selected work

Sr. #	Scientific database	Type	Selected research works	No. of researches
1	IEEE	Journal	[6, 7]	2
		Conference	[8–12]	5
2	Springer	Journal	[13, 14]	2
		Conference	[15–19]	5
3	Elsevier	Journal	[3, 20–23]	5
		Conference	[4, 24]	2
4	ACM	Journal	[25]	1
		Conference	[1, 26–30]	6
5	Taylor and Francis	Journal	[31–33]	3
		Conference	–	–

Table 5 Identification of user categories

Sr. #	User identification categories	Percent (%)	References
1	Behavioral identification	45	[3, 6, 7, 11, 13, 20–25, 28, 30–32]
2	Attribute identification	32	[4, 8 9, 11 15, 16, 23, 29, 32, 33]
3	Topic identification	12	[10, 26, 27, 29]
4	Spam identification	16	[1, 12, 14, 18, 19]
5	Crime identification	12	[4, 16, 17, 23]

to have received assertive research attention in many years. There are some studies in the literature that contains two or more user identification categories in one study.

3.3 User Categorization Models and Algorithms (RQ 2)

Considering the data extracted from the answer to this research question, it emerges that behavioral, attribute, spam and crime identification models and algorithms are cited by both journal and conference papers while topic identification frameworks are just mentioned in conference papers.

Different models with the goal of identifying the user, its behavioral patterns and attributes have been listed in this research question. The studies reporting the identification of user behaviors through short text were [11, 21]. In publications [7, 13, 28, 32] user's interest and influence regarding responsiveness along with communication and exploration were identified. Demographic features of user like age, gender, education, date and email address were identified by models cited in [4, 16, 33]. Spam emails in [14] were spotted by using the anti-spam model as shown in Table 6.

Table 6 User categorization models

Sr. #	User categorization models	References
1	**Behavioral(12)** Comment tree model, UR, UC and UCR model, UCT, DSUN, DSM, pipe-lined system models, AS-LDA and LDA multi-layer perception model, SVM classification model, personalized recommendation model, DTM, research models	[7, 8, 11, 13, 20, 21, 23–25, 28, 30, 32]
2	**Attribute(7)** Machine learning models, pipe-lined system models, SVM classification model, classification models, authorship identification model, research models, theoretical models	[4, 11, 15, 16, 23, 32, 33]
3	**Topic(2)** Hidden Markov model, vector space model	[26, 27]
4	**Spam(1)** Anti-spam model	[14]
5	**Crime(2)** Pipe-lined system models, authorship identification model	[16, 23]

Table 7 User categorization algorithms/methods

Sr. #	User categorization algorithms/methods	References
1	**Behavioral(6)** Measurement method, gibbs sampling algorithm, IITP improved semi-supervised algorithm, sentiment flow algorithm Generation process algorithm	[6, 11, 13, 21, 23, 24, 28, 30]
2	**Attribute(5)** ReLU, IR and ML algorithm, IITP, stylometry method, improved semi-supervised algorithm	[4, 9, 11, 23, 29]
3	**Topic(3)** Feature term method, IR and ML algorithm group recommendation method	[10, 26, 29]
4	**Spam(3)** Random forest, real-time (DeBOT) method, ML and compression algorithm	[12, 18, 19]
5	**Crime(2)** Grammar derivation, grammar combination and general EFG algorithm, IITP	[22, 23]

After the detailed analysis of 31 studies, we have distinguished 10 algorithms and 5 methods that have been trained through machine learning techniques to make specific decisions as demonstrated in Table 7.

Table 8 User categorization tools

Sr. #	User identification tools	Purpose	References
1	Third party tools	URLs recognition	[1]
2	Knowledge based tools	Emotion recognition	[3]
3	ROUGE	Automatic text summarization	[27]
4	LIWC tool	Personality/behavior recognition	[9, 28]
5	ITAP	Behavior recognition	[23]
6	Stanford POS tagger	Read text and assign POS	[8]
7	Automated tools	Spam recognition	[12]
8	Word segmentation tool	Word recognition	[15]

Gibbs sampling algorithm which is used for feature selection in researches [21, 24, 30] is the only algorithm used in multiple papers. Algorithms can be applied on supervised, unsupervised and semi-supervised learning depending upon the dataset. In research [11] a semi-supervised learning algorithm for semi labeled data was used to train the data until it is labeled completely. Most of the algorithms in this research study were used on supervised learning. Random forest [12] and machine learning and compression [19] algorithms were used for classification and regression.

3.4 User Categorization Tools (RQ3)

This section of the study presents the tools which are used in 9 research papers as shown in Table 8 to act as user identification. The basic purpose of these tools is to reduce the ambiguity in the text present on different social network platforms.

Tools used in this research question belong to both research community and public sector. Multiple third-party tools [1] like Browsing API, SURBL, and Spamhaus from both research and private sectors and automated tools [12] are detecting malicious URLs and spam tweets. Knowledge-based tools [3] was developed in contrast with statistical approaches to analyze and extract knowledge from each sentence to specify its sentiment status. There are some tools that automatically evaluate documents, ROUGE [27] is used to evaluate generated summaries with the summaries created by experts. LIWC [9, 28] is the only tool in the research which is used by two studies to recognize behavioral patterns. The University of Austin builds IITP tool to describe the criminal process, vulnerabilities, and resources that facilitate criminals to commit the crime. Two natural languages processing tools POS Tagger [8] and Word Segmentation tool [15] were used to extract features from tweets and recognize words respectively.

4 Discussion and Limitation

In this study, we evaluated and identified text mining techniques that help and support to distinguish users, demographic features of users and behavioral patterns while communicating on different social networking sites. The data used in this research is mostly gathered from Twitter. Out of 31 studies, 13 used twitter data for their experimentation. Other data sources include Facebook, Blogs, documents, reports and instant messages.

In text classification, machine learning techniques vary for different datasets. Different text requires a different set of features and ML techniques. Preprocessing, a data mining technique transforms data into an understandable format, was used in 14 papers mostly where natural language processing is performed on the text, to attain optimal achievement. Correct feature selection increases the accuracy and performance of the classifier. Most frequently used feature selection techniques were POS Tagging and TF-IDF, use of these techniques improved the performance of the machine learning models. The finding suggests that future studies adopt both semantic based features and demographic features together to achieve higher performance.

Classifiers are performed on supervised learning to validate the experimental results. In this research classifiers are used by 14 studies and support vector machine (SVM) alone is used in 9 studies separately as well as combined with other classifiers. Selected studies suggested that performance of SVM in text classification is much better than other classifiers like random forest and naïve Bayes.

For user categorization, we identified 19 models, 10 algorithms, 8 tools and 5 methods. In some studies like [11, 13, 23, 26, 28, 30] they are used together to improve the performance. It has been shown in this review; frameworks perform differently on every dataset depending upon the size and type of text used in datasets. Therefore, before making any decision on the choice of models, algorithms and ML techniques, professionals not just should know about the performance, yet also need to comprehend the qualities of the frameworks.

Table 9 shows the comparison of text mining techniques (models, algorithms/methods, and tools) that have been proposed for user identification based on the type of datasets, pre-processing, feature selection, classifier, and validation. It has been observed that mostly algorithms/methods are validated against state of the art machine learning techniques as compared to models and tools that have been identified. Whereas, the comparison based on pre-processing also shows that most of the algorithms/methods used pre-processing step as compared to tools and models.

In this review for accessing the performance of text mining techniques, only accuracy metrics is observed. If a model or algorithms fail to perform below the minimum threshold in terms of accuracy practitioners will reject it, although in addition to accuracy metrics other evaluation metrics such as propagation ability and accountability is ignored in this review can also necessarily be considered. Table 9 shows the discussion and comparison of all 31 papers selected in this literature.

Table 9 Comparison of text mining techniques

Research	Datasets	Pre-processing	Feature selection	Classifier used	Tools	Models	Algorithms/methods	Validation	References
R1	Facebook				✓			✓	[1]
R2	ISEAR	✓	BOW (Bag of words)	Ensemble and Naïve Bayes	✓				[3]
R3	Twitter	✓	POS tagging	SVM and random forest		✓	✓		[4]
R4	Text document	✓	TF and term weight	SVM and Naïve Bayes		✓	✓	✓	[26]
R5	Blogs and news reports	✓			✓	✓			[27]
R6	Reddit	✓	TF-IDF		✓	✓	✓		[28]
R7	Last.fm			Ensemble and SVM		✓		✓	[25]
R8	Twitter and blogs	✓	POS tagging	SVM and Naïve Bayes			✓		[29]
R9	Twitter					✓	✓		[30]
R10	Twitter					✓			[20]
R11	Twitter	✓	TF-IDF	K-means		✓	✓	✓	[21]
R12	WITS	✓	N-grams				✓	✓	[22]
R13	News and reports	✓	POS tagging and named entity recognition		✓	✓	✓	✓	[23]
R14	Corpus	✓	LDA	SVM		✓	✓	✓	[24]
R15	Twitter	✓	POS tagging	K-means and EM	✓	✓		✓	[8]
R16	Chat room conversations		Chi-square	SVM and Naïve Bayes	✓		✓		[9]

(continued)

Table 9 (continued)

Research	Datasets	Pre-processing	Feature selection	Classifier used	Tools	Models	Algorithms/methods	Validation	References
R17	SinaWeibo						√		[10]
R18	SinaWeibo	√		SVM		√	√		[11]
R19	Blog data	√	SenticNet and POS tagging	Ensemble			√		[6]
R20	Yelp					√			[7]
R21	Twitter		ReLF	Random forest	√		√	√	[12]
R22	Instant messages		POS tagging and feature based selection	SVM and neural networks	√	√			[15]
R23	SinaWeibo		Density based selection			√	√	√	[13]
R24	Email and blogs		TF-IDF	SVM		√			[16]
R25	Twitter, email and IM	√	Stylometry						[17]
R26	Facebook and Twitter						√		[18]
R27	Twitter	√	BOW (Bag of words)	SVM and Naïve Bayes			√	√	[19]
R28	Facebook and Twitter		Content based feature						[31]
R29	Facebook and Twitter					√		√	[32]
R30	Facebook and Twitter					√			[33]
R31	Email					√			[14]

5 Conclusion and Future Work

In this literature review, we have provided empirical evidence on the existence of a mapping between identities of individuals across the social media sites and studied the possibility of identifying users across sites. Both link and content information was used to identify users. A Systematic Literature Review (SLR) has been used to determine 31 studies published during 2010–2018. Identifying users across social media sites opens the door to many interesting applications such as analyzing usage patterns across networks and studying user behavior. We have identified five different user identification categories and corresponding algorithms, methods, models, and tools. Identifying different user behaviors has the potential to improve business and resource management in OSN's.

For future direction, we could investigate, for instance, recommendation systems that exploit the user behaviors to display more appropriate advertisements. We could exploit the user behaviors to define different classes and develop more accurate performance models for the service.

References

1. Gao H, Hu J, Wilson C, Li Z, Chen Y, Zhao BY (2010) Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Nov 2010. ACM, pp 35–47
2. Tuna T, Akbas E, Aksoy A, Canbaz MA, Karabiyik U, Gonen B, Aygun R (2016) User characterization for online social networks. Soc Netw Anal Mining 6(1):104
3. Perikos I, Hatzilygeroudis I (2016) Recognizing emotions in text using ensemble of classifiers. Eng Appl Artif Intell 51:191–201
4. Sboev A, Litvinova T, Gudovskikh D, Rybka R, Moloshnikov I (2016) Machine learning models of text categorization by author gender using topic-independent features. Proc Comput Sci 101:135–142
5. Kitchenham B (2004) Procedures for performing systematic reviews, Keele, UK, Keele University, vol 33, no 2004, pp 1–26
6. Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A (2015) Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Comput Intell Mag 10(4):26–36
7. Qian X, Feng H, Zhao G, Mei T (2014) Personalized recommendation combining user interest and social circle. IEEE Trans Knowl Data Eng 26(7):1763–1777
8. Murkute AM, Gadge J (2015) Framework for user identification using writeprint approach. In: 2015 international conference on technologies for sustainable development (ICTSD), Feb. IEEE, pp 1–5
9. Amuchi F, Al-Nemrat A, Alazab M, Layton R (2012) Identifying cyber predators through forensic authorship analysis of chat logs. In: 2012 third cybercrime and trustworthy computing workshop (CTC), Oct. IEEE, pp 28–37
10. Wang J, Liu Z, Zhao H (2014) Group recommendation using topic identification in social networks. In: 2014 sixth international conference on intelligent human-machine systems and cybernetics (IHMSC), vol 1, Aug. IEEE, pp 355–358
11. Yin C, Xiang J, Zhang H, Wang J, Yin Z, Kim JU (2015) A new SVM method for short text classification based on semi-supervised learning. In: 2015 4th international conference on advanced information technology and sensor application (AITS), Aug. IEEE, pp 100–103

12. Meda C, Ragusa E, Gianoglio C, Zunino R, Ottaviano A, Scillia E, Surlinelli R (2016) Spam detection of Twitter traffic: a framework based on random forests and non-uniform feature sampling. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), Aug. IEEE, pp 811–817
13. Guo H, Chen Y (2016) User interest detecting by text mining technology for microblog platform. Arab J Sci Eng 41(8):3177–3186
14. Zhang Y, He J, Xu J (2018) A new anti-spam model based on e-mail address concealment technique. Wuhan Univ J Nat Sci 23(1):79–83
15. Ding Y, Meng X, Chai G, Tang Y (2011) User identification for instant messages. In: Neural information processing. Springer Berlin/Heidelberg, pp 113–120
16. Ma J, Teng G, Chang S, Zhang X, Xiao K (2011) Social network analysis based on authorship identification for cybercrime investigation. Intell Secur Inf 27–35
17. Frommholz I, Al-Khateeb HM, Potthast M, Ghasem Z, Shukla M, Short E (2016) On textual analysis and machine learning for cyberstalking detection. Datenbank-Spektrum 16(2):127–135
18. Chavoshi N, Hamooni H, Mueen A (2016) Identifying correlated bots in twitter. In: International Conference on Social Informatics, Nov. Springer International Publishing, pp 14–21
19. Santos I, Minambres-Marcos I, Laorden C, Galán-García P, Santamaría-Ibirika A, Bringas PG (2014) Twitter content-based spam filtering. In: International joint conference SOCO'13-CISIS'13-ICEUTE'13. Springer, Cham, pp 449–458
20. Zhou X, Wu B, Jin Q (2017) User role identification based on social behavior and networking analysis for information dissemination. Future Gener Comput Syst
21. Qiu Z, Shen H (2017) User clustering in a dynamic social network topic model for short text streams. Inf Sci 414:102–116
22. Sharef NM, Martin T (2015) Evolving fuzzy grammar for crime texts categorization. Appl Soft Comput 28:175–187
23. Zaeem RN, Manoharan M, Yang Y, Barber KS (2017) Modeling and analysis of identity threat behaviors through text mining of identity theft stories. Comput Secur 65:50–63
24. Liang J, Liu P, Tan J, Bai S (2014) Sentiment classification based on AS-LDA model. Proc Comput Sci 31:511–516
25. Chelmis C, Prasanna VK (2013) Social link prediction in online social tagging systems. ACM Trans Inf Syst (TOIS) 31(4):20
26. Manne S, Fatima SS (2012) An extensive empirical study of feature terms selection for text summarization and categorization. In: Proceedings of the second international conference on computational science, engineering and information technology, Oct. ACM, pp 606–613
27. Chakraborti S (2015) Multi-document text summarization for competitor intelligence: a methodology based on topic identification and artificial bee colony optimization. In: Proceedings of the 30th annual ACM symposium on applied computing, Apr. ACM, pp 1110–1111
28. Choi D, Han J, Chung T, Ahn YY, Chun BG, Kwon TT (2015) Characterizing conversation patterns in Reddit: from the perspectives of content properties and user participation behaviors. In: Proceedings of the 2015 ACM on conference on online social networks, Nov. ACM, pp 233–243
29. Inches G, Crestani F (2011) Online conversation mining for author characterization and topic identification. In: Proceedings of the 4th workshop on workshop for Ph.D. students in information & knowledge management, Oct. ACM, pp 19–26
30. Zhao Y, Liang S, Ren Z, Ma J, Yilmaz E, de Rijke M (2016) Explainable user clustering in short text streams. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, July. ACM, pp 155–164
31. O'Riordan S, Feller J, Nagle T (2016) A categorisation framework for a feature-level analysis of social network sites. J Decis Syst 25(3):244–262
32. Son JE, Lee SH, Cho EY, Kim HW (2016) Examining online citizenship behaviours in social network sites: a social capital perspective. Behav Inf Technol 35(9):730–747
33. Riedl C, Köbler F, Goswami S, Krcmar H (2013) Tweeting to feel connected: a model for social connectedness in online social networks. Int J Hum-Comput Interact 29(10):670–687

Automated Analysis of Eye-Tracker-Based Human-Human Interaction Studies

Timothy Callemein, Kristof Van Beeck, Geert Brône and Toon Goedemé

Abstract Mobile eye-tracking systems have been available for about a decade now and are becoming increasingly popular in different fields of application, including marketing, sociology, usability studies and linguistics. While the user-friendliness and ergonomics of the hardware are developing at a rapid pace, the software for the analysis of mobile eye-tracking data in some points still lacks robustness and functionality. With this paper, we investigate which state-of-the-art computer vision algorithms may be used to automate the post-analysis of mobile eye-tracking data. For the case study in this paper, we focus on mobile eye-tracker recordings made during human-human face-to-face interactions. We compared two recent publicly available frameworks (YOLOv2 and OpenPose) to relate the gaze location generated by the eye-tracker to the head and hands visible in the scene camera data. In this paper we will show that the use of this single pipeline framework provides robust results, which are both more accurate and faster than previous work in the field. Moreover, our approach does not rely on manual interventions during this process.

Keywords Post analysis · Human-Human-interaction studies
Mobile eye-trackers

T. Callemein (✉) · K. Van Beeck · T. Goedemé
KU Leuven - EAVISE, Jan Pieter de Nayerlaan 5, Sint-Katelijne-Waver, Belgium
e-mail: timothy.callemein@kuleuven.be

K. Van Beeck
e-mail: kristof.vanbeeck@kuleuven.be

T. Goedemé
e-mail: toon.goedeme@kuleuven.be

G. Brône
KU Leuven - MIDI, Sint-Andriesstraat 2, Antwerp, Belgium
e-mail: geert.brone@kuleuven.be

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_50

1 Introduction

A growing field of application for mobile eye-trackers is the recording of human-human interactions, enabling researchers in the fields of linguistics and conversation analysis to analyse the role of eye gaze in non-verbal communication and interaction management. This research, among others, focuses on the distribution of gaze of each interlocutor during face-to-face interactions, answering basic research questions such as: “How long does a person spend looking at the face or hands of an interlocutor during a conversation?”, “Does the distribution of visual attention differ depending on the type of interaction, the role or status of the participants, or other factors?” Mobile eye-trackers contain the necessary hardware to simultaneously record the scene from the wearer’s perspective and track the gaze of the wearer during this recording, providing an insider’s perspective on the interaction.

Most software currently provided with mobile eye-trackers comes with a basic Area-of-Interest (AOI) analysis method that allows the user to select a bounding box as AOI, for example for determining specific objects of interest for an analysis of visual attention. Afterwards the software matches the chosen AOI to the current gaze location using the matching technique discussed in [1]. This technique provides an automatic annotation of rigid objects containing similar features as the model. During human-human interactions, however, the main AOI would be faces of co-participants or hands performing gestures, but unfortunately both are non-rigid and impossible to recognize based on simple appearance-based techniques, such as in [1]. The lack of an automatic annotation tool will leave most research in this field resorting to manual annotation of these types of ‘objects’. Depending on the amount of people taking part in the study, the amount of data grows, resulting in a cumbersome time-consuming annotation task. In this paper we investigate if state-of-the-art computer vision techniques can perform accurate detection of hands and heads as the most relevant non-rigid objects for human interaction analysis, without making use of artificial markers. These detections will, in a second step, be combined with the gaze coordinates of the eye-tracking camera, producing a fully automatic annotation tool which will eliminate a significant part of the manual annotation work. As an illustration, Fig. 1 shows a frame from a three-person-conversation (from the perspective of one of the co-participants wearing eye-tracking glasses) with (a) the output of a pose estimation algorithm [2] and (b) a red circle representing the wearer’s current gaze fixation.

The remainder of this paper is organized as follows. In Sect. 2 we discuss related work on previous post-analysis techniques with mobile eye-trackers, followed by three sections explaining our approach on detecting the torso in Sect. 3. Based on the results of this initial detection step, we will advance to Sect. 4, limiting the detection area to the head. Section 5 focuses on the hands using the same techniques as explained in the previous sections. In Sect. 6 we discuss the results comparing previous work with both the YOLOv2 detector and pose estimation based detections. The paper closes off with a general conclusion and suggestions for future work in Sect. 7.

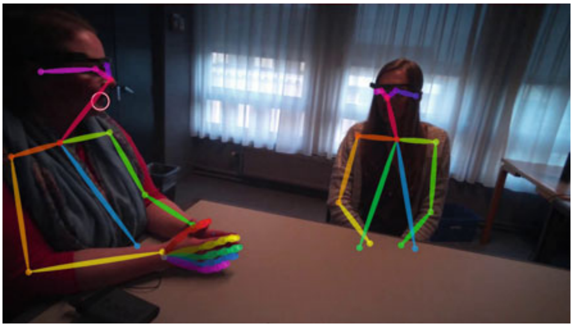


Fig. 1 Frame from a three-persons-conversation processed by the pose estimator (coloured skeletons) with the gaze (red circle)

2 Related Work

Most eye-tracking systems are shipped with manufacturer software assisting the user in automated annotation and data aggregation, next to providing a manual annotation tool for more fine-grained analysis. Mobile eye-trackers, while providing an advantage in mobility, come with the disadvantage of a dynamic spontaneous scene. The most advanced manufacturers ship software based on appearance-based image features (SIFT [3], SURF [4]) providing a rudimentary single-image-model annotation applicable for studies focussing on rigid objects, but not be able to deal with non-rigid three-dimensional objects like hands and heads that change during the course of time and vary from person to person.

The work of De Beugher et al. [5] tries to answer the need for an automated annotation tool capable of annotating non-rigid objects by using a machine learned model on multiple images. They use two techniques, one to train a more generalised model of a person's upper-body based on a deformable part model (DPM) [6], while the second is trained to detect faces based on HAAR features [7]. Both of these techniques are evaluated separately and combined in [5] on an annotated dataset. For the purpose of this paper we will focus on the performance of the detector, leading to better results during gaze classification. Furthermore, the upper-body comprises not only of the head location, offering no guarantee that the gaze placed upon the upper-body can be classified as the head due to its larger AOI.

Apart from the head we want to automate the annotation process for hands that appear in the scene camera images generated by the eye-tracker, providing a first basic coding layer for e.g. gesture studies. A hand detector combining a two stage hypothesis and classification method by [8] shows that a single model can be improved by taking other properties of our hands into account. De Beugher et al. improved this work further in [9] by adding their previous upper-body model as a pre-processing step. Additionally, they implemented a rotating hand model and allowed manual interventions to further improve accuracy. Yet, this model still suffers from the different orientations of the hand opposed to the orientation of the trained model. Furthermore, they only achieve good results when sufficient manual interventions are given

(1.61% gives an F1-score above 85.17%). In our research we develop fully-automatic detectors with no obtrusive elements and no need for manual interventions.

Previously mentioned techniques involving machine learning seem insufficient to yield high accuracy in detecting a difficult object like the human hand. These techniques are recently being overtaken by state-of-the-art neural network object detectors capable of extracting high levels of features that comprise an object. Although deep learning solutions have been kept in the background for quite a while, recent advancements in General Purpose Graphics Processing Unit (GPGPU) hardware and an increase in available training data (e.g. *ImageNet* [10]) have allowed their emergence. The current deep learning techniques greatly outperform previous handcrafted and simple machine learning techniques. In this paper we compare two state-of-the-art deep learning techniques and test their accuracy for the automatic annotation of mobile eye-tracking data.

3 Torso Detection

One of the baselines of human face-to-face interaction is the simple fact that interlocutors tend to gaze at the face of the other while interacting (with addressees typically gazing at the current speaker more and longer than the other way around [11, 12]). This makes the head one of the prime objects to be detected as part of an automated annotation procedure.

Previous techniques tried to detect the head location by using the upper body or torso including the head [5]. This allows the detector to use more information, leading to a better result. In this paper we compare two state-of-the-art deep learning based techniques with traditional upper-body detectors. In our research we have focussed on using the state-of-the-art YOLOv2 detector [13] based on the Darknet framework for retraining purposes. We first annotated the torsos of 4000 images from the dataset provided by [8]. We then used this data by including pre-trained weights on the VOC-dataset [14] to calculate new weights to detect the torso.

The second technique that we included, which is also based on deep learning, is called pose estimation. Pose estimators, compared to a conventional detector, will not only produce a bounding box around the person or detection in general. They try to estimate the separate key-points of body-part-joints that together compose the pose of that person. Using the key-points that are part of the torso, we can use the pose estimator as a torso detector, by returning a bounding box around them. In this paper we have implemented the OpenPose framework [15] bundling three components. The first part is capable of detecting the separate anatomic body joint points (e.g. shoulder, elbows, wrists, …). When only one person is visible all the found points will belong to that person. When there are multiple people in the image, however, the body joint points will have to be grouped according to the person they belong to. The second part includes a network capable of detecting the part affinity fields (PAF) between joints [2]. These PAF will assist the previous network in combining the joint points to the corresponding person. The last part consists of detecting a more detailed pose of the hands, which will be discussed further in Sect. 5.

4 Head Detection

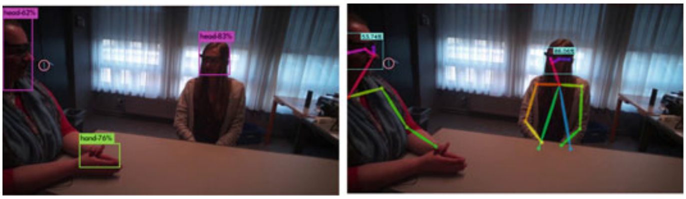
In the previous section we used two state-of-the-art techniques in order to detect the torso. As in [5], this can be used to acquire context information on where to find the head. Moreover, if one is only interested in the question whether the test person looks at another person, a torso detector suffices. However, as mentioned above, researchers are generally interested in the question if the gaze is directed towards the face or head of the interlocutor.

In Sect. 3, we described how we retrained the YOLOv2 detector on a specific dataset in order to train a torso detector. The retraining was based on around 1800 annotations from the manually annotated dataset used in [16], containing both the head location and the hand key-points.

We have followed the same approach as with the torso-model to train a YOLOv2 model on this data, but included both the head and hand into a single detector.

Using the pose key-points we were able to find the torso. However, these points are also usable to estimate the head position. Yet the pose estimator only provides the eyes, ears and nose point. We therefore determined a bounding box around the head based on these points by first looking at the head direction with respect to the camera. When the head is frontal we return a bounding box based on the ear-by-ear distance and nose point. However, when the head is in profile this is not possible. Not only the centre will shift away from the nose, but some points will be self-occluded by the head. By detecting the orientation of the head beforehand (e.g. frontal, right profile, left profile), we can provide a more accurate bounding box around the head.

Due to the margin of error presented by current mobile eye-trackers it is possible that the gaze cursor is focussed on the head, yet is not within the strict boundaries of the head. This error margin may increase over time, especially with longer recordings with a single calibration step before the start. During manual annotation, the final decision will depend on experience with the eye-tracker and general offset present on the eye-tracker. Figure 2 illustrates both the pose-based and YOLOv2-model based detections of the head with the gaze near the head, but not within the boundaries. In our model we have included the option to increase the margins. This allows for a bigger head boundary and a consistent annotation decision process. However, in this paper during the evaluation of our detections in Sect. 6 we used stricter boundaries to not influence our detection results.



(a) YOLOv2 based head/hand detection

(b) OpenPose based head detection

Fig. 2 Both techniques detecting the head

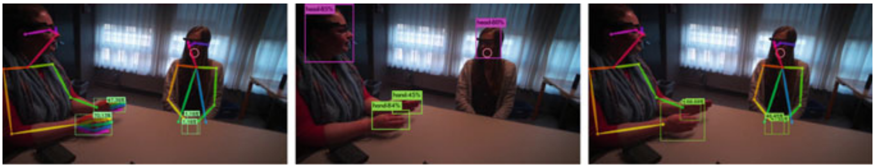
5 Hand Pose Estimation

Apart from the head, the hands also play a central role in non-verbal communication as prime articulators of visible bodily action. We therefore compare different state-of-the-art detectors that may be useful as part of the annotation of eye-tracking data. A particular challenge here is that fast motion of both the hands to be detected, as well as movements of the head by the person wearing the eye-tracking glasses may result in the hands being blurred and unclear in the images to be processed. This contributes to the difficulty level of detecting them in an accurate way. In Sect. 4 we trained a combined YOLOv2 model.

Another hand pose estimator was presented by [16]. They use the wrist location from the complete pose as a basis during the hand-pose estimation. This model is capable of estimating each separate hand joint separately.

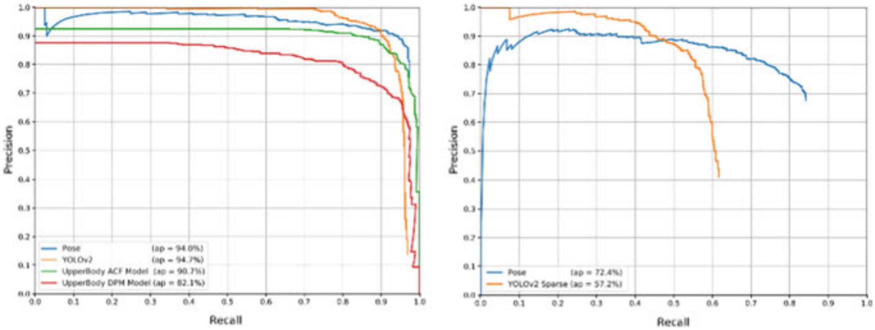
In order for the hand pose estimator to work, enough detail of the hand must be visible. When the image is unclear or blurred we notice that the estimator fails. We therefore developed an additional pose-based hand detector by using pose points of the arm. A bounding box is estimated around the hand based on the length and direction of the vector between the wrist and elbow.

Figure 3 illustrates the pose estimated hand, the hand pose detection and the YOLOv2 hand-head model detection. Our initial intention was to combine the pose estimated hand detection with the hand pose detection, yet they seemed to contradict each other. The estimated detection will be present even if the hands are occluded, as illustrated in Fig. 3c, while Fig. 3a and b show no detection or a very low detection confidence. Both situations can be favourable depending on the aspect of the study and thus are complementary.



(a) Hand pose based detection *(b) YOLOv2 Hand detection* *(c) Hand estimated based on the elbow-wrist*

Fig. 3 Three methods compared for hand detection



(a) *Torso comparison on Inria* (b) *Head comparison on Inria*

Fig. 4 YoloV2 and OpenPose comparison on Inria

6 Results

6.1 Torso

Our first technique concerns the torso detections compared to non-deep learning techniques. In the work of De Beugher et al. [17], two state-of-the-art upper body models were tested on the INRIA person dataset [18] containing person annotations. The dataset only contained full person detections, which was compensated for in [17] by only taking into account the top 66% of the person as upper body. We evaluated our torso YOLOv2 model and pose-based torso detector on the same data. Our results in the form of PR-curves are illustrated in Fig. 4a together with the UpperBody aggregated channel features (ACF) [19] and UpperBody DPM [6] models from [17]. These precision and recall curves were generated by varying a threshold over the detection confidence scores.

For the pose-based technique we have calculated the mean of all separate joint confidence scores. Both the pose-based and YOLOv2 technique show an increased average precision opposed to the ACF and DPM model.

Table 1 F1-scores on the InsightOut [9] and 5-Signers [20] datasets

	Mittal et al. [8]	Yang and Ramanan [21]	De Beugher et al. [9]		Ours			
				Incl. tracking	Pose estimated sparse dense			
D1 (%)	85	24.2	83.4	88.2	98.4	97.6	92	**99.4**
D2 (%)	48.9	46.5	52.9	65.3	**91.1**	84.8	48.2	61.9
5-signers (%)	77.6	n.a.	81.1	n.a.	**97.6**	88.3	84	92.2

6.2 Head

Because only detecting the torso does not suffice as a basis for annotation, as mentioned above, we evaluated our head approach on the same INRIA dataset. The INRIA dataset contained the head location and the person bounding box. We used 66% of the person annotation width as a reference for the head annotation size. Figure 4b illustrates the precision and recall of the YOLOv2 and pose-based model. The recall drops faster compared to the results in Fig. 4a, which is to be expected since the head opposed to the full torso has a smaller area, increasing the detection challenge with a high confidence score. Here the pose-based detection clearly outperforms the YOLOv2 model, mainly because the pose-based detections have the full yet hidden pose as decision support. The YOLOv2 model has no such support and has to rely on the head only.

6.3 Hands

To evaluate the hand approaches we used the InsightOut dataset (D1 and D2).

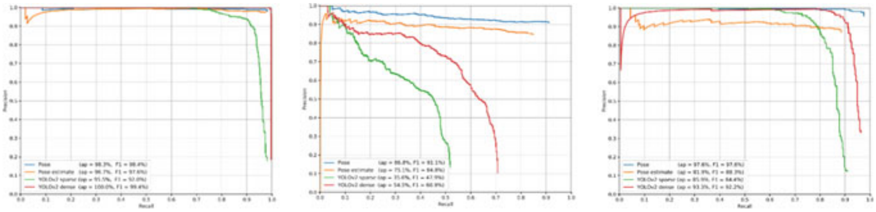
De Beugher et al. [9] and the 5-Signers dataset [20]. Table 1 compares the F1-scores of the different approaches with our work. Our results show that our proposed approaches all show an increase in F1-score opposed to existing models. Only the YOLOv2 (*Sparse, Dense*) models show a slight decrease in accuracy on the D2-dataset compared to the work of [9].

Besides the accuracy, we also compared the processing times of our techniques (Table 2). Previous techniques only used the CPU to process the data, whereas our approaches require a mid-end GPU (NVIDIA GTX 1080 Ti) capable of running the used algorithms. We conclude that only the YOLOv2-based approach is able to run in real-time, although the proposed pose-based techniques are at least 70 times faster than the competitors.

To compare these techniques against each other we plotted the PR-curve for each approach on each dataset, illustrated in Fig. 5a–c. Comparing YOLOv2 against the pose-based techniques shows that both YOLOv2 models are less accurate on the D2

Table 2 Execution time comparison

	Mittal et al. [8]	Yang and Ramanan [21]	De Beugher et al. [9]	Ours		
				Pose	Estimated	YOLOv2 models
Avg time/frame (s)	293.33	113	36.67	0.5	0.125	**0.0099**
Avg fps	0.00341	0.00885	0.02750	2	8	**100**



(a) Hand results on YOLOV2 and OpenPose on D1

(b) Hand results on YOLOV2 and OpenPose on D2

(c) Hand results on YOLOV2 and OpenPose on 5-Signers

Fig. 5 Results on three hand datasets

dataset. We see that training a denser model on more hand data increases the mAP of the model compared to the sparse two class model.

As expected we observe that the pose-based techniques produce good results. The estimated hand location based on the elbow and wrist shows a decreased precision, which is expected due to the static direction on which we estimate the hand. In reality the hand does not necessarily follow the arm movement explaining the performance drop compared to the hand pose estimator.

6.4 Automated Annotations

The main goal of our work, providing automatic annotations during human-human interactions, will produce labels on the recordings, depending on overlap between the detections and the gaze. When the gaze falls within the boundaries of a detection, the detection label will be the coded for that frame. To evaluate the head gaze labels, we manually annotated 2500 frames of a face-to-face spontaneous conversation between three people wearing mobile eye-trackers. On this segment we used both the OpenPose head detector and YOLOv2 sparse head model to generate automatic labels for each frame. In case of manual annotation of recordings it may occur that opinions differ between annotators.

To overcome this source of discussion the annotations are commonly compared by different measures to obtain a score of resemblance (referred to as an intercoder agreement test or ICA-test) In order to compare our automated labels with the ground truth we use the same tests to obtain scores evaluating our techniques. These results

Table 3 Reliability levels of the automated head annotation

	OpenPose level (%)	YOLOv2 level (%)
Agreement [22]	91.1	**92.6**
Scott's pi [23]	82.2	**85.3**
Cohen's kappa [24]	82.3	**85.3**
Krippendorf's alpha [25]	82.2	**85.3**

are visible in Table 3. The annotated video of this sequence, using the Pose based head and hand detection can be viewed on https://youtu.be/eEVXIfY99O0.

7 Conclusion

This paper focussed on comparing the current state-of-the-art techniques on automated mobile eye-tracking analysis. This involves detecting the hands and heads appearing in the scene images generated by the eye-tracker, which may be obvious foci of attention during face-to-face human interactions. The output of this automatic detection step provides a solid basis for further annotation of relevant non-verbal behaviour, including hand gestures, head movements and so on. We compared two main techniques, a pose estimator and the YOLOv2 detector against more traditionally used techniques showing an overall higher accuracy. Despite the fact that these deep learning techniques require a mid-end GPU, they easily achieve faster than real-time performance. By providing multiple techniques and allowing adjustable bounding boxes margins, the gaze annotations are customizable according to the requirements of the specific study at hand without any need for manual intervention.

References

1. Brône G, Oben B, Goedemé T (2011) Towards a more effective method for analysing mobile eye-tracking data: integrating gaze data with object recognition algorithms. In: Proceedings of the PETMEI. ACM, pp 53–56
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017)
3. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. IJCV 60(2):91–110
4. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. ECCV 2006:404–417
5. De Beugher S, Brône G, Goedemé T (2014) Automatic analysis of in-the-wild mobile eye-tracking experiments using object, face and person detection. In: VISAPP, vol 1. IEEE, pp 625–633
6. Felzenszwalb PF, Girshick RB, McAllester D (2010) Cascade object detection with deformable part models. In: CVPR. IEEE, pp 2241–2248
7. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: CVPR, vol 1. IEEE, pp I–I

8. Mittal A, Zisserman A, Torr PH (2011) Hand detection using multiple proposals. In: BMVC, pp 1–11
9. De Beugher S, Brône G, Goedemé T (2015) Semi-automatic hand detection: a case study on real life mobile eye-tracker data. In: Proceedings VISAPP 2015, vol 2. SciTePress, pp 121–129
10. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 248–255
11. Oertel C, Wlodarczak, M., Edlund J, Wagner P, Gustafson J (2012) Gaze patterns in turn-taking. In: Thirteenth annual conference of the international speech communication association
12. Brône G, Oben B, Jehoul A, Vranjes J, Feyaerts K (2017) Eye gaze and viewpoint in multimodal interaction management. Cogn Linguist 28(3):449–483
13. Redmon J, Farhadi A Yolo9000: better, faster, stronger
14. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2009) The PASCAL visual object classes challenge 2009 (VOC2009) results
15. Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: CVPR
16. Simon T, Joo H, Matthews I, Sheikh Y (2017) Hand keypoint detection in single images using multiview bootstrapping. In: CVPR
17. De Beugher S (2016) Computer vision techniques for automatic analysis of mobile eye-tracking data. PhD thesis, KU Leuven, Belgium
18. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 1. IEEE, pp 886–893
19. Yang B, Yan J, Lei Z, Li SZ (2014) Aggregate channel features for multi-view face detection. In: IJCB 2014. IEEE, pp 1–8
20. Buehler P, Everingham M, Huttenlocher DP, Zisserman A (2008) Long term arm and hand tracking for continuous sign language tv broadcasts. In: BMVC, pp 1105–1114
21. Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. IEEE, pp 1385–1392
22. Neuendorf KA (2016) The content analysis guidebook. Sage
23. Scott WA (1955) Reliability of content analysis: the case of nominal scale coding. Public opinion quarterly, pp 321–325
24. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46
25. Krippendorff K (2012) Content analysis: an introduction to its methodology. SAGE

Part VII
Software Engineering

Survey on Compromise-Defensive System Design

Tomas Cerny and Michael Jeff Donahoo

Abstract Conventional enterprise application design methodologies emphasize performance, scalability, and development/maintenance costs. Often such applications deal with access to confidential data (e-commerce, health, etc.). A single flaw in the application may lead to a compromise, exposing computational resources and sensitive data, such as private information, trade secrets, etc. Traditionally, security for enterprise applications focused on prevention; however, recent experience demonstrates that exploitation of infrastructure, operating systems, libraries, frameworks, personnel, etc. are almost unavoidable. While prevention should certainly remain the first line of defense, system architects must also incorporate designs to enable breach containment and response. In this paper, we survey related research on software application design that targets isolation, where the compromise of a single module presents a knowable and scope-limited worst-case impact.

Keywords Compromise · Isolation · Security · Modularization · Survey

1 Introduction

The consequences resulting from software application compromise can be significant. The future application design could limit data exposure, which would, in turn, limit the impact of a potential compromise. While lots of existing works address compromise prevention, a single error on the defending side may result in system compromise. When that happens, it is often too late to mitigate the damage. Only a limited research emphasis has been given to prevention of data leaks from compromised applications. For operating systems and cloud computing, efforts are notable

T. Cerny (✉) · M. J. Donahoo
Baylor University, One Bear Place #97141, Waco, TX 76798, USA
e-mail: tomas_cerny@baylor.edu

M. J. Donahoo
e-mail: jeff_donahoo@baylor.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_51

in application-level isolation, but the isolation elements were not designed to mitigate the compromise impact.

This paper considers existing techniques and works that address the impact mitigation of compromises. This survey focuses on information and content management systems. Thus, it excludes research related to application sandboxing or multi-application environments that is more fitting to anti-virus tools; instead, it considers distributed, possibly cloud-based, microservice or SOA-like applications and their design. It begins with providing a roadmap to existing research. In the next section, it introduces the context of existing technologies. Next, it looks into existing evidence available in IEEE Xplore, ACM Digital Library, and Springer Link. The conclusion highlights a design practice that would mitigate the consequences resulting from application compromise.

2 Background

The primary focus in conventional design emphasizes two aspects: development and scale. Typically, software engineers deal with development and maintenance efforts by logically dividing the functionality into various modules, while emphasizing "low coupling" among module interaction and "high cohesion" of individual modules. Textbooks [1] provide broad details on conventional decomposition and modularization, highlighting the evolved best practices.

With respect to scale, individual modules are deployed to distinct servers, bringing more computational resources to bear [2] and providing module distribution and balancing. Benefits include better fail-tolerance with multiple modules instances handling higher request loads. The evolution of distributed systems evolved over time from Common Object Request Broker Architecture and Java RMI, settling on the Service-Oriented Architecture (SOA) decomposition, and introducing an Enterprise Service Bus (ESB) for service integration, with a service orchestration enforcing various business or security policies.

From today's perspective, SOA is becoming obsolete [2] and the industry is implementing new systems through another type of modularization called Microservice Architecture (MSA). It is a cloud-friendly approach emphasizing low coupling among modules, maintaining its own scheme of the database with bounded-context and, thereby, reducing dependencies across modules. This further brings a measure of independence to modules and prevents error propagation across module evolution. MSA modules are usually deployed through containers that isolate modules, primarily for the purpose of enabling on-demand module replication/initiation/relocation when the request load grows to support elasticity. Each module instance is given limited CPU/Memory resources, which accommodate the module needs. On the other hand, two different-purpose MSA modules may both maintain the same information, forcing the design to introduce replication, negatively impacting development and maintenance.

Security concerns rarely play a major role when it comes to application modularization. Designers prefer to focus on securing systems from outside compromise; they typically do not prepare the modules to mitigate the losses in case a breach happens. In most approaches, user-machine communication goes through Secure Socket Layer (SSL), which dedicates a particular module for request balancing and routing and another for user management, providing a Single-Sign-On Service (SSO). User identity is supplied to all independent modules through SSO. From the perspective of authorization, modules (especially MSA modules) usually deal with authorization on their own, deteriorating global policy governance/enforcement in the enterprise throughout the evolution. Moreover, there are various levels of security enforcement tangling through the entire application, deteriorating performance with additional detail lookups (e.g., user details, user associations) and extending development efforts.

Although current security mechanisms aim to prevent intrusion from the application interface, there are various places where the application can get compromised. The attacks even include a compromise of the underlying operating system and the physical machine the application uses. An application can also be attacked from the inside. For instance, an operating system may apply updates from a compromised or fake update server,[1] leading into a full compromise of the system. The application may use an unsecured web server [3]. An administrator's account can be compromised due to fishing, human error, etc. There are simply way too many intrusion points in current systems. To make the situation worse, there is always the human factor, increasing the risk due to mistakes, typological errors, etc. Never-ending system updates, evolution, server updates, operating system updates, etc., requires the overall systems to be constantly adapted and managed, leaving a potential for an intrusion. When such an intrusion happens, it may lead to catastrophic consequences.

3 Existing Knowledge

To establish the base of existing knowledge in this field, we researched publications at three indexing sites, IEEE Xplore, ACM Digital Library and SpringerLink, using a keywords search, similar to a systematic mapping study [4]. The used search terms were selected to identify a compromise as the topic, as well as isolation topics. However, to narrow the scope of the papers, the selection is restricted to works related to either containers, or microservices, since industry defines it as the future architecture for cloud-based applications [2]. Alternatively, we used as the limiting factor the topics of the scope limitation. To extract the query, we considered a control set from works we considered related to the aimed area. The resulting query consists of terms "compromise" and "isolation", with refined results to the occurrence of any of the terms "container", "scope" (meaning "limiting scope"), or "microservice". This produces the following query:

[1] Ukraine cyber-attack: Servers seized, 2017, http://www.bbc.com/news/technology-40497026.

```
(compromise AND isolation) AND (container OR scope OR microservice)
```

From this search, we evaluated 492 papers based on abstracts, titles, and if matched on full text. We extracted 207 publications from IEEE Xplore, 112 from ACM Digital Library, and 173 from SpringerLink. Furthermore, we spot checked evidence at the less credible but open and voluminous Google Scholar.

The considered area of compromise-defensive system design has been addressed by multiple researchers in different perspectives. In particular, the perspective of *isolation* is well described by [5]. It suggests that one application's failure must not adversely affect unrelated applications or the system itself. From our aimed perspective, this applies when considering module isolation, whether or not it is within a centralized or distributed system.

Often *virtualization* is used to consolidate servers or infrastructure for efficient resource utilization as well as for isolation [6]. There are two major virtualization approaches, *virtual machines* and *containers* providing operating-system-level virtualization. Containers are widely believed to outperform virtual machines because of small virtualization overheads, while virtual machines are expected to provide stronger performance isolation.

When dealing with data access in cloud systems, researchers [7] have looked into the Platform as a Service (PaaS) model coping with failover and disaster recovery. The platform provides virtualization techniques like virtual machine managers/hypervisors and containers. The researchers evaluated the security benefits of both containers and hypervisors, specifically, the Hyper-V and Docker containers. The results suggested that hypervisors lack a sufficient logical isolation configuration to prevent a side channel attack; the security flaws among multi-tenants allowed breaches of the confidentiality and integrity of data. Containers are much more secure and provide much stronger isolation at the micro level, and they overcome the issues that we can find with hypervisors. Applications can consider PaaS and containers as they bring isolation mechanisms.

Utilization of modularization principles such as MSA [2] in containers could divide a complex system into multiple self-contained modules in terms of functionality and data storage. In contrast to other system modularizations, such as SOA, where all data could end up in a single schema, MSA contains relevant process-flow, limited data perspective, constraints, and internal know-how. MSA is exposing as few-as-possible, as opposed to SOA, delegating the process-flow and possibly business logic on the integration level. Each microservice can evolve, and deploy, independently of other services, as long as it does not violate expected interaction. On-demand service scaling is the main goal for MSA, where multiple, identical services hide behind a balancer. The tax for such level of service isolation and independence is data and source code replication across multiple distinct services. The MSA approach is promising for the direction and purpose of isolation-based design.

Content-based isolation made progress in modern platforms with isolation policies [8]. From the per-user perspective isolating users but letting applications running within the same isolation container, recently evolved a need for application isolation policy. Different applications are isolated from one another. This introduces novel challenges, such as mutually-distrusting content elements interfering with one another inside a single application. For example, compromised photo editor app, which happens through a harmful image, allows the attacker to steal other images processed by the editor. The content-based isolation approach can prevent that kind of theft. However, to support a flexible isolation granularity and preserve system compatibility remains a challenge.

Simultaneous execution of multiple software components on top of the same virtual machine may break down any isolation between them [9]. Single component compromise can span to the whole system and consume all available resources. In [9], authors identify faulty software components and survey monitoring solutions for detecting anomalies. They suggest to instrument suspicious components for evaluation and leave others to avoid performance degradation.

Communication across the isolation barrier in enterprise apps deployed in isolated environments is addressed in [10]. The aim is to prevent intrusion or fault and show how to design inter-application interaction support in Operating System (OS)-level virtualization systems without causing significant compromise of the virtual machine isolation. In a prototype, authors can preserve isolation capability for inter-application interactions in a variety of sample applications.

The challenge of OS component-level isolation to avoid a compromise in one kernel's module escalating into a compromise of the entire kernel is discussed in [11]. In particular, authors advocate disk encryption involving two-factor authentication for decryption, so as to prevent a compromise in one kernel's module from compromising the whole machine. This established knowledge could be transferred to modularization in MSA-based approach.

To avoid compromise spreading in cloud computing, the compromised host must be isolated and subsequently recovered. The work [12] presents multiple complex strategies to isolate the host, stop impacted services, deactivate compromised accounts, and disinfect, delete or quarantine compromised files.

To prevent data leakage from compromised systems or insider attacks from curious administrators an encryption can be applied to data storage [13]. The system can decrypt data access only with the knowledge of individual's personal password. As a result, the administrator never gets access to decrypted data, and even if all servers are compromised, an adversary cannot decrypt the data of any user who is not logged in. However, traffic monitoring in a compromised system may reveal user passwords, their hashes, or the decrypted data. One limitation is access to data needed for reports and statistics. In addition, not all data storage and querying operations are available or provide expected performance. Even with access to shared data, it may not be possible to perform aggregations. On the other hand, a hybrid schema could be adopted placing sensitive and private data in an encrypted vault limiting the impact from a system compromise.

Addressing data leaks from compromised web applications is researched in [14]. Specifically, authors aimed to prevent bulk data leaks caused by code injection in web applications as well as compromised user-level processes on the application server. User login was associated with each user session, enforcing information-flow tracking. The system flagged each file and database record to ensure that application data is released only to the sessions of authorized users. Extra focus was placed on isolation of data between user sessions. This approach fits well with banking and e-commerce applications. The most common server-side web app attacks could be addressed by the approach. The performance overhead was about 20%–30% over unmodified applications. Similarly, in [15], the authors let each web browser application run in its own virtual machine to provide isolation from each other and from the user's resources.

Processing sensitive data in healthcare or finance must comply with legal policies regarding their release. Software bugs may lead to an irreversible disclosure of confidential data. In [16], the authors question how to guarantee the release of sensitive data only to authorized users in enterprise applications. They considered an event-based application backend to be a middleware for processing units and a broker that acts as a "safety net" that prevents a harmful data disclosure. It associates data with security labels, similar to [14], and transparently tracks their propagation at different granularity across the system architecture with storage and complex event processing through the frontend. Their approach is used in the UK National Health Service (NHS).

Distributed systems often utilize a dedicated module for authentication, providing SSO service across various modules. The authentication module can issue labels similarly to [16]. In service orchestration, the authorization label could be requested from all services involved in the composition to perform the composite service. Moreover, a similar balancing module could perform hashing/routing to the appropriate module, when horizontal module-granularity exists (modules operate with data fragmented horizontally, e.g., "A–F", "G–L", etc.).

A programming model for security reviews in web applications is introduced in [17]. It enables verification of high-level security properties by dividing the application into isolated components and limiting the privileges allotted to each one. The researchers consider session isolation similarly to the one described in [14] as well as component isolation in a way that components cannot tamper each other. Such a model should provide robust security at the component level since it restricts what each component of the application can do and quarantines compromised code. It provides a way to safely integrate third-party, less-trusted code into a web application. However, the application must be designed in a custom language in order to enforce isolation components. This is a radically new approach for application development; a special environment is introduced to run the system, and special enforcement is made to code components to enforce isolation (no mutable state allowed). Next, the application is considered to be a monolith. When evaluated, this model brought a 20% throughput penalty.

To determine the extent of the damage caused by compromise [18] target the OS running the applications. An application-level isolation and recovery system was developed to limit the effects of the compromise while simplifying the post-intrusion recovery process. The work used a copy-on-write file system with restricted privilege isolation environment for running untrusted applications. The explicit file sharing mechanism across the isolation environments limits attack propagation without compromising functionality. If a sandboxed application proves to be untrustworthy, a coarse-grained recovery method allows removing the footprint of the software.

Usage of an untrusted environment when working with sensitive data is addressed in [19]. It is similar to distributed systems with a compromised module. While standard approaches for sensitive data protection are too strict or unreliable [19] proposes a virtualization-based approach for preventing sensitive data leaks from a computer running an untrusted commodity OS. Their security system is realized by the hypervisor that concurrently runs two virtual machines under the control of the identical untrusted OS. The first (private) VM is used for sensitive data processing and has no network access. The second (public) VM has Internet access and is used solely as a system call server. Hypervisor intercepts system calls executed by the explicitly selected trusted applications in the private VM and forwards them for execution through a tamper-proof memory channel into the public VM. The hypervisor protects memory and control transfer integrity of the trusted processes from unauthorized changes made by the malicious OS kernel. It ensures that the potentially-compromised OS is unable to access the channel by injecting malicious code into the process address space or transferring control to it without being caught by the hypervisor.

As can be seen from existing work, when it comes to data, encryption can be applied [13]; however, this limits database operations and does not help with resource theft or exposure of know-how from in compromised modules. We can also monitor data [16] by issuing labels, but that approach is limited only to data. Various isolation approaches already exist to address scaling issues, recovery [7], content-based isolation [8], application isolation [9], etc., but they have not been used for the purpose of compromise isolation. These existing techniques and technologies can be employed to help design a modularization architecture designed specifically for isolating compromises. Certain required capabilities have been already studied in different contexts [10–12, 17] and may be adapted to design compromise-defensive systems.

4 Conclusion

This paper provides an overview of existing work on isolation-based application design approaches facing the rising threats from system compromise. When we could limit the compromise impact in our application to a measurable scope, a particular security breach will not lead to devastating cases losing third-party or sensitive data; our infrastructure will not turn into bitcoin miners, and the internal know-how theft becomes limited. The finer granularity of isolation leads to more limited impact. The tradeoff goes towards development efforts, system operations, and performance. Having an application divided into multiple MSA that either address different functionality or a selected set of users seems like two categories on how an application can isolate.

Illustrating existing works on the topic, we show existing efforts addressing data privacy, either through their tracking, labeling or encryption. Application isolation considers virtual machines and containers. This also applies for MSA targeting scalability and system operation simplicity. When dealing with distributed systems sharing data, we must assume that one of these could be compromises and thus multi-component communication is addressed by the number of works. Multiple works deal with application sandboxing to determine whether an application is compromised or not, these usually reach low-level and OS.

In future work, we will address the isolation-based design with Aspect-Oriented Transformation taking a conventional or even monolith application extended with additional markup (join points) as an input, transforming it into multi-module distributed application following two patterns; one considering different functionality, and the other considering distinct user sets. MSA seems as the approach for isolation-based systems, we aim to utilize it. Our secondary goal will be to find the optimal distribution and size of modules for a particular, available computational environment (hardware).

References

1. Larman C (2004) Applying UML and patterns: an introduction to object-oriented analysis and design and iterative development, 3rd edn. Prentice Hall PTR, Upper Saddle River
2. Cerny T, Donahoo MJ, Trnka M (2018) Contextual understanding of microservice architecture: current and future directions. SIGAPP Appl Comput Rev 17(4):29–45
3. Spring T (2017) 3.2 million servers vulnerable to jboss attack (2017). https://threatpost.com/3-2-million-servers-vulnerable-to-jboss-attack/117465/
4. Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering. Inf Softw Technol 64(C):1–18
5. Back G, Hsieh WC (2005) The kaffeos java runtime system. ACM Trans Program Lang Syst 27(4):583–630
6. Mardan AAA, Kono K (2016) Containers or hypervisors: which is better for database consolidation? In: 2016 IEEE international conference on cloud computing technology and science (CloudCom), pp 564–571
7. Samo JA, Ahmed Z, Shaikh A (2017) Advocating isolation of resources among multi-tenants by containerization in IaaS cloud model. In: Innovations in electrical engineering and computational technologies (ICIEECT), pp 1–17
8. Moshchuk A, Wang HJ, Liu Y (2013) Content-based isolation: rethinking isolation policy design on client systems. In ACM SIGSAC conference on computer & communications security. CCS '13, NY, USA, ACM, pp 1167–1180
9. Gonzalez-Herrera I, Bourcier J, Daubert E, Rudametkin W, Barais O, Fouquet F, Jzquel J (2014) Scapegoat: an adaptive monitoring framework for component-based systems. In: IEEE/IFIP conference on software architecture, pp 67–76
10. Shan Z, Wang X, Chiueh Tc, Meng X (2012) Facilitating inter-application interactions for os-level virtualization. SIGPLAN Not 47(7):75–86
11. Richter L, G¨otzfried J, Muller T (2016) Isolating operating system components with intel SGX. In: Proceedings of the 1st workshop on system software for trusted execution. SysTEX '16, New York, NY, USA, ACM, pp 8:1–8:6
12. Taheri MA, Jaatun MG (20112) Handling compromised components in an IaaS cloud installation. J Cloud Comput: Adv Syst Appl 1(1):16

13. Popa RA, Redfield CMS, Zeldovich N, Balakrishnan H (2011) Cryptdb: protecting confidentiality with encrypted query processing. In: Proceedings of the twenty-third ACM symposium on operating systems principles. SOSP '11, New York, NY, USA, ACM, pp 85–100
14. Mundada Y, Ramachandran A, Feamster N (2013) Silverline: preventing data leaks from compromised web applications. In: The 29th annual computer security applications conference. ACSAC '13, New York, NY, USA, ACM, pp 329–338
15. Cox RS, Gribble SD, Levy HM, Hansen JG (2006) A safety-oriented platform for web applications. In: Proceedings of the 2006 IEEE symposium on security and privacy. SP '06, Washington, DC, USA, IEEE Computer Society, pp 350–364
16. Hosek P, Migliavacca M, Papagiannis I, Eyers DM, Evans D, Shand B, Bacon J, Pietzuch P (2011) SafeWeb: a middleware for securing ruby-based web applications. Springer, Berlin, pp 491–511
17. Krishnamurthy A, Mettler A, Wagner D (2010) Fine-grained privilege separation for web applications. In: Proceedings of the 19th international conference on World Wide Web. WWW '10, New York, NY, USA, ACM, pp 551–560
18. Jain S, Shafique F, Djeric V, Goel A (2008) Application-level isolation and recovery with solitude. In: Proceedings of the 3rd ACM SIGOPS/EuroSys European conference on computer systems 2008. Eurosys '08, NY, USA, ACM, pp 95–107
19. Burdonov I, Kosachev A, Iakovenko P (2009) Virtualization-based separation of privilege: Working with sensitive data in untrusted environment. In: Proceedings of the 1st EuroSys workshop on virtualization technology for dependable systems. VDTS '09, New York, NY, USA, ACM, pp 1–6

A Study on Cost-Effective and Eco-friendly Bicycle Sharing System for Developing Countries

Larsson Bajracharya, Tirta Mulya, Ayi Purbasari and Mintae Hwang

Abstract As part of the development of information technologies for eco-friendly transportation system, we performed a study on cost-effective bicycle sharing system feasible for developing countries. The sharing system will work based on online registration and real-time monitoring system, using smart phone and passcode. A low budget Kiosk system for bicycle access from the station will be used. This provides efficiency to users as well as the system for smooth and easy operation.

Keywords Bicycle-share · Cost-effective · IoT · Passcode

1 Introduction

The current digital age has helped us to apply technological reforms in various day to day activities, and it is thus necessary for the application of technology to create a greener, environment-friendly reforms in transportation infrastructures. However, even when implementing changes in transport-based rules and vehicle management system, it is impossible to bring a striking change when it comes to green energy

L. Bajracharya (✉)
Department of Eco-Friendly Offshore Plant FEED Engineering, Changwon National University, Changwon, Republic of Korea
e-mail: larssonbajra@gmail.com

T. Mulya · A. Purbasari
Department of Informatics Engineering, Pasundan University, Bandung, Indonesia
e-mail: tirta.mulia@unpas.ac.id

A. Purbasari
e-mail: pbasari@unpas.ac.id

M. Hwang
Department of Information & Communication Engineering, Changwon National University, Changwon, Republic of Korea
e-mail: professorhwang@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_52

and eco-friendly transport system. Therefore, a shared bicycle-system may have a potential for significant change when it comes to addressing this matter.

Bicycle sharing system is a rental system that is usually established in a commercial basis, which allows users to rent a bicycle from specific station and return it to any rental station, within the same city. The purpose is to provide public with cheap access of bicycle for a short period of time, as an alternative to motorized transportation [1].

The following guidelines are applicable to efficiently share bicycles [2]. First, try to calculate the city size, population density and topography and climate of a city, which helps in efficient placement of stations. Second, consider the technology being used for the system as well as capital and operating cost. Third, try maintaining a feasible security system for the bicycles and stations, as well as a status monitoring system. This paper contributes to study on a cost-effective globally feasible bicycle sharing system, that can optimize the use of public bicycles to maintain an eco-friendly environment.

The paper is organized as follows. Chapter 2 introduces a related research on bicycle sharing system as well as IoT technology, Chap. 3 briefly introduces the main idea on overview of bicycle-sharing system, and Chap. 4 presents the challenges that the system might face in a developing country's real environment. Finally, Chap. 5 presents the conclusion of this paper.

2 Related Research

We are all familiar with growing population of the world, eventually increasing the use of transport facilities. Increasing number of these vehicles have seen a lot of problems in the cities like air and noise pollution. Moreover, narrow roads and unplanned city transportation system has made the problems even severe. This is making bicycle use more important than ever before.

Study on Internet of Things states it as a paradigm that is rapidly gaining ground around in the scenario of modern communication. The idea is to bring together various things around us which through unique addressing schemes, will be able to interact with each other [3]. This paradigm sees the application in different fields, such as home automation, industrial areas, medical assistance, elderly assistance, traffic management and many others. The main idea of an IoT system is to create a smarter way of life using computing tasks instead of everyday mechanical tasks.

Another study analyzing the IoT based system [4] discussed the urban IoT architecture being used, based on web servers, various link layer architectures and the devices that are essential to realize an urban IoT. It also discussed on backend server and database management for IoT based systems.

A study on sustainable mobility in urban areas [2] discussed an optimized network to anticipate asymmetric travel demands of large cities. For example, people would love to come down a hill with a bicycle but would not climb up with one. Therefore, dedicated vehicles for redistribution of bicycles are necessary. But even though dedicated vehicles for bicycle redistribution is possible, large number of redistribution would cause latency and inconvenience. Therefore, discount for bicycles being returned to stations with higher elevation can be introduced.

Also, a study on social and environmental sustainability [5] discussed socio-cultural practices that play a vital part in bicycle sharing system. During this study, various key experts that were affiliated with bicycle sharing systems around Europe were interviewed, helping us to get an in-depth knowledge on political and social practices, as well as systemic and legislative changes necessary for a successful bicycle share system.

Moreover, study on bicycle share strategies and environmental sustainability [6] show current measures being used for bicycle security and management. Despite physical damages done by the users, it is important to know how securely bicycles can be left in a certain area. Instead of a physical locking system, Kiosk- based system [7] seems to be popular and highly secure.

3 Basic Architecture of the Bicycle-Sharing System

Research and implementation of various bicycle sharing systems are currently present at large, aiming to resolve various socio-economic as well as environmental problems. The architecture of this system holds similarity in features and performance with any other system, but light weight technology and cost effectiveness will make it feasible for developing countries. As in Fig. 1, we can view the architecture of this system.

The user and bicycle related data will be stored and managed in the server database, and it is thus possible to search anytime and anywhere using a mobile-based application. As shown in the figure, there will be a two-way communication between the server and application, as well as the bicycles present in the docks of the station via kiosk control box. The communication will help to know the whereabouts of bicycles and the position of users.

In this paper, a mobile and IoT based bicycle-share system is proposed to target public transport users. The whole system can be mainly separated into two categories, i.e. mobile application and station/server system.

3.1 Mobile Application System

As our main idea is based on a mobile based sharing, an application for a handheld device is proposed. The application will run in a suitable operating system environment, helping the user to get the passcode for desired bicycle he/she wishes to rent. But before the bicycle is rented, the application verifies whether the user is valid to rent, i.e. he/she should have an account to login with an id and password as well as valid date in his account. Validity date can be extended by recharging the

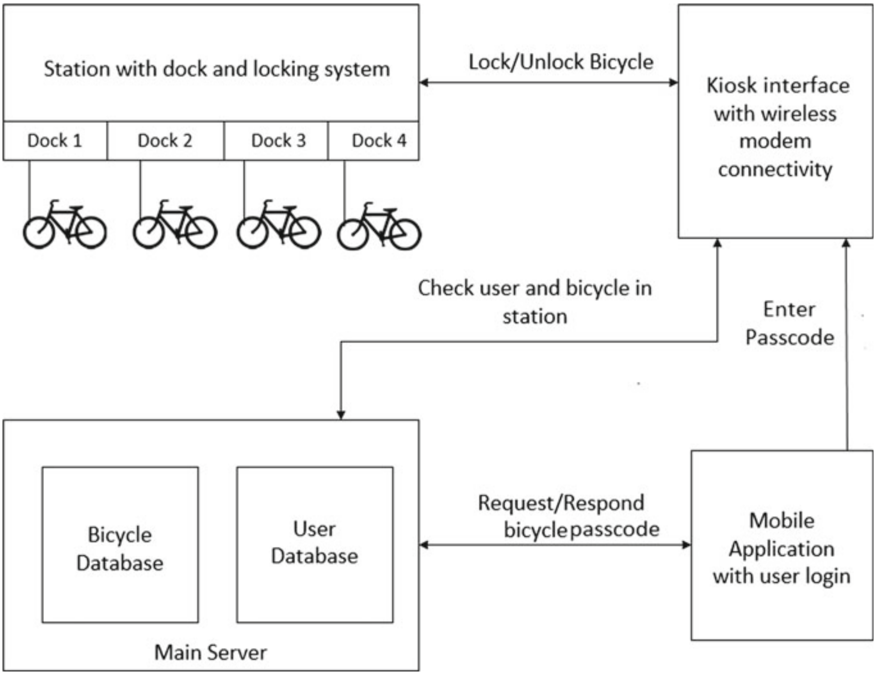
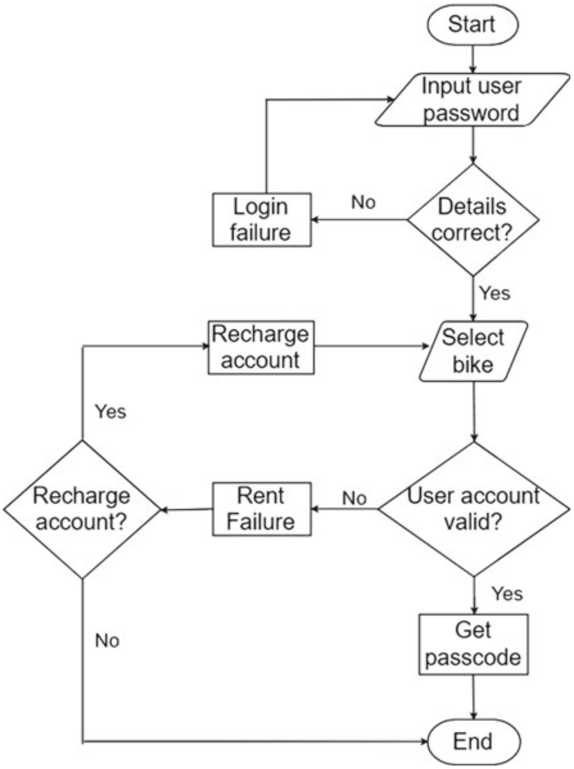

Station with dock and locking system
Dock 1
Dock 2
Dock 3
Dock 4
Lock/Unlock Bicycle
Kiosk interface with wireless modem connectivity
Check user and bicycle in station
Enter Passcode
Bicycle Database
User Database
Main Server
Request/Respond bicyclepasscode
Mobile Application with user login

Fig. 1 Simple architecture of bicycle sharing system

account. Figure 2 below shows a flowchart of the implementation idea of a mobile-based bicycle-share application.

The application will have multiple services targeted towards the user for safe and easy riding experience. Users will be able to book desired bicycles online irrespective of place or time. He/she will receive a passcode that can be used to unlock as well as park the bicycle in any desired station with a free dock. The booked bicycle will be reserved for the specific user for 15–20 min without anyone else being able to access it. If he/she fails to unlock the bicycle from the station, the passcode received earlier will be invalid and an alert message stating that the time period has been expired will be sent and the bicycle will be available for other users to book. As account recharge for online system is always a hurdle, the system will therefore be able to accept any online payment methods, i.e. domestic or international cards using a payment gateway or a simple bank transfer selecting the desired bank from the application itself. Integration of google maps in the application will help the user to search the nearest station for bicycle rent, parking or choosing a feasible route. To keep track of one’s rental records or details of total distance covered, a graphical table with charts will be presented to the customers. In case of any emergency, the customer will be able to search for the nearest mechanic, hospital or police station using the emergency features available in the application. This application will also help the bicycle system to create its own community in social media. Communication

Fig. 2 Flowchart of a mobile-based bicycle-share application



between fellow bicycle enthusiasts, social events as well as any discomfort regarding the bicycle service or any station around the city can be discussed to improve the system. Figure 3 shows the various discussed services provided by the application.

3.2 Station/Server System

Server for the system will be maintained in a remote location along with a periodic back up system. A database management system (MSSQL or MySQL) will be used to maintain the system data. The server will also consist of a wireless network system (3G or 4G) to communicate with the remote stations as well as the users. Likewise, each station around the city will consist of a low budget Kiosk system for bicycle rental or parking, developed using various IoT based components like Arduino, Raspberry Pi, GPS locator, motors, LCDs, touch pad etc. The Kiosk system in each station needs to run a specific integrated set of programs so that it can receive and transmit the instructions, as well as perform the required task, as requested by the server. Each of these programs running will be different or will be modified with

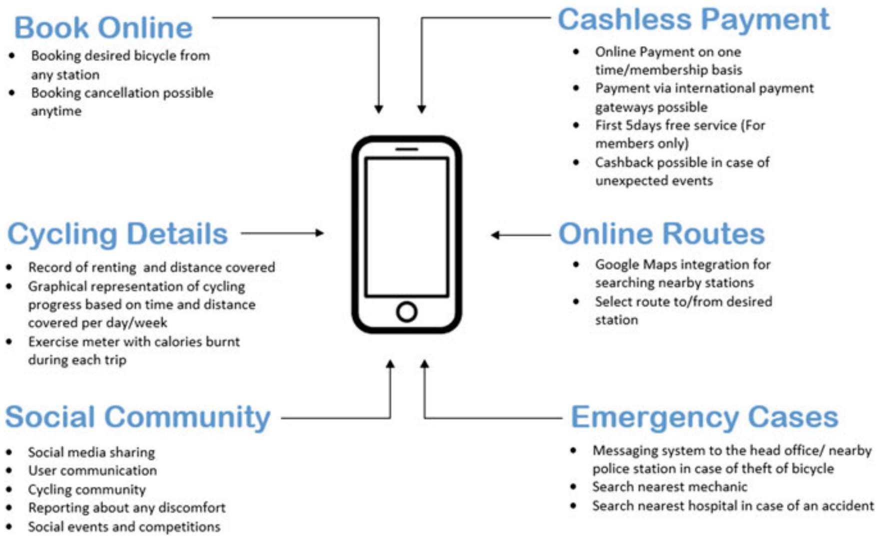

Book Online
Booking desired bicycle from any station
Booking cancellation possible anytime
Cashless Payment
Online Payment on one time/membership basis
Payment via international payment gateways possible
First 5days free service (For members only)
Cashback possible in case of unexpected events
Cycling Details
Record of renting and distance covered
Graphical representation of cycling progress based on time and distance covered per day/week
Exercise meter with calories burnt during each trip
Online Routes
Google Maps integration for searching nearby stations
Select route to/from desired station
Social Community
Social media sharing
User communication
Cycling community
Reporting about any discomfort
Social events and competitions
Emergency Cases
Messaging system to the head office/ nearby police station in case of theft of bicycle
Search nearest mechanic
Search nearest hospital in case of an accident

Fig. 3 Various services provided by the application

time, depending upon the number of bicycle docks added. For data communication, 4G network can be used where available, whereas GSM or 3G network can also be used where 4G services are not yet available. To make the system eco-friendly, solar panels will be maintained in each station, along with a backup battery to power up the control box (Kiosk) along with all the other hardware components. Figure 4 below shows a flowchart of the data processing in each station, after the user requests using the Kiosk system interface and Fig. 5 shows a detailed workflow of the system numbered from 1 to 14, starting from user request for passcode from the servers.

4 Challenges for the Proposed System

As our system mostly targets developing countries, issues like budget, topography, social awareness about use of new technology are major challenges. Moreover, using a wireless communication interface for data transmission holds higher threat in these countries, as the network security and stability is still in its infant stage. Also, when we use IoT sensors and actuators, it is highly possible that hardware failure or corruption occurs. Some of the major challenges of the proposed system are:

- Lack of knowledge about technology
  Most of the developing countries are not familiar with use of information technology for commercial purpose. Lack of technological advancement and being used to perform most of the tasks in traditional manner may cause people to think twice before using an application and IoT based system.

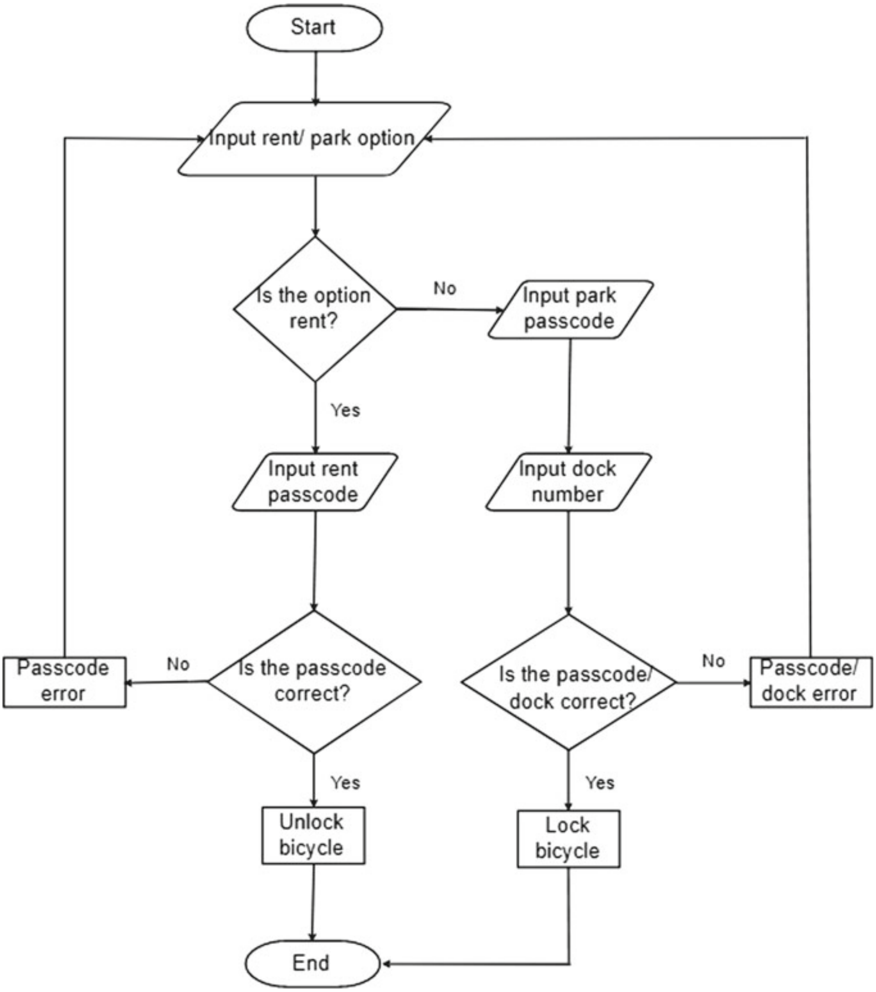


Fig. 4 Flowchart of data flow and processing in the station

- Budget for the system
  A public bicycle sharing system might be bit of a hurdle for governments in developing countries to implement soon. Management of various stations, bicycles, regular maintenance as well as automated pay system for transportation is a challenge. Moreover, a proper bicycle lane is also required everywhere around the city.
- Topography of the city
  As every city around the world has different topography, it is challenging to well plan the network of bicycle stations around. Also, implementation of IoT based sharing means network needs to be maintained actively all the time. Topography

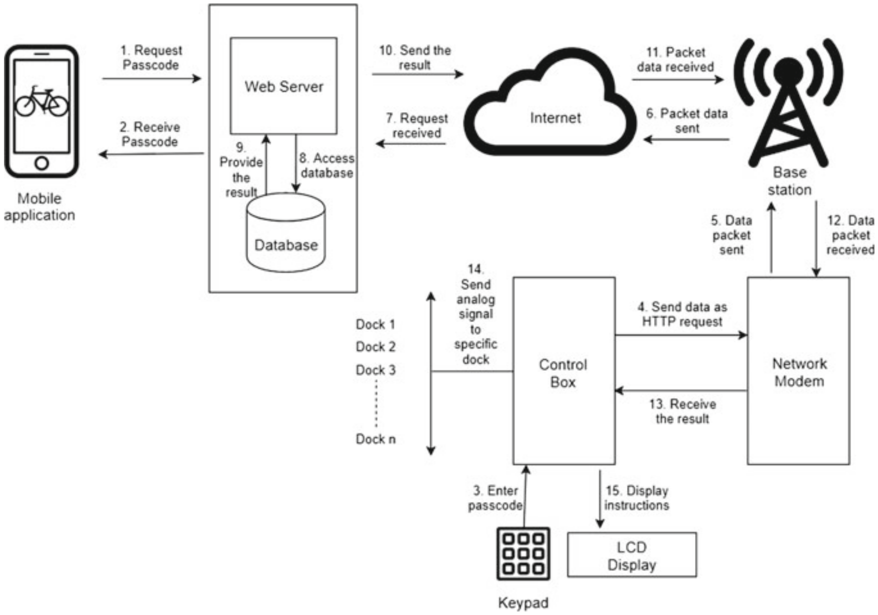


Fig. 5 Detailed workflow of the bicycle sharing system

and lack of better resources to maintain stable network may be a huge challenge to overcome. Moreover, poor traffic management plays a major setback to implement the system. Therefore, study on GIS (Geographic Information System) is necessary for better overall result.

- Data Corruption
  Sometimes the attacker may try to change the transmitted data. This is done by transmitting valid frequencies of data spectrum at correct time. Time can be calculated if attacker has good knowledge of modulation scheme and coding. As data security in developing countries are low, the system needs to develop a data encryption program to prevent threat from the intruders.

5 Conclusion

The population today is growing every day and the city transportation system is not being able to meet the demands of public. To tackle these problems and make the city eco-friendly, we studied about a Cost-Effective Bicycle Sharing System implementing IoT and mobile application. We discussed that use of technology like IoT and application-based sharing system is possible in cities across the world with lower budget, so that people travel more on bicycles and use less public or private means of transportation. We also discussed various challenges this system might face which might make it difficult for developing countries. Therefore, discussion for an effective way of implementing good telecommunication and IoT based system in those countries are very necessary.

IoT being used in developing countries in bicycle sharing might lead to overall change in conventional lifestyle as well. Bicycle sharing systems being used by other countries have seen a lot of positive impact in their society. Even though a lot of challenges need to be resolved for this system, but an economically feasible bicycle sharing system may provide a solution to many of the problems faced today in cities around the world concerned with environment and means of transportation. In future, introducing system like NFC cards for payment and lock/unlock verification system will further help to enhance efficiency and secure transaction.

Acknowledgements This paper was supported by Changwon National University Research Fund 2017.

References

1. Marchuk M, Shkompletova A, Boyarskaya A (2016) Bicycle sharing system. In: 72nd student scientific and technical conference, pp 339–342
2. Midgley P (2011) Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. Comm Sustain Dev 9(10):5–12
3. Atzori L, Lera A, Morabito G (2010) The internet of things: a survey, pp 1–3
4. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities 1(1):25–28
5. Figueroa M, Greiffenberg C, Akinyi E, Brink A, Behrendt C, Krlev G, Navratil J, Placier K, Cavola F, Turini A, Cancellieri G (2016) Social innovation in environmental sustainability 6(4):26–55
6. www.onbikeshare.com (2016) Bike share implementation strategies: a comparative guide, pp 3–7
7. Holfelder W, Hehmann D (1994) A networked multimedia retrival management system for distributed kiosk applications

An Approach to Modeling Microservice Solutions

Zhiyi Ma, Jinyang Liu and Xiao He

Abstract Modeling occupies an important place in microservice solutions. However, as far as approaches covering whole development cycle are concerned, either their modeling languages are too simple, or modeling processes are incomplete. This paper presents an approach to modeling microservice solutions based on CBDI SAE metamodel for SOA 3, which has gotten the widely attention of academic and industrial circles. The paper discusses which modeling activities can output which models and how to build and describe the models, and prescribes the relations between the models.

Keywords Modeling · Microservice · Metamodel

1 Introduction

Microservice solutions have become a research hotspot at present, and modeling occupies an important place in the solutions.

An approach to modeling microservice solutions should include a modeling language and a modeling process guideline. As far as current approaches covering whole

Z. Ma (✉)
School of Electronics Engineering & Computer Science, Peking University, Beijing, China
e-mail: mazhiyi@pku.edu.cn

Z. Ma
Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing, China

J. Liu
Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
e-mail: 1601214751@pku.edu.cn

X. He
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
e-mail: hexiao@ustb.edu.cn

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_53

development cycle are concerned, either their modeling languages are too simple, or modeling processes only refer to one or several development phases. Fowler only gives eight principles for building microservices, and does not discuss who to model microservices [1]. Mazzara et al. provide four programming concepts for addressing microservice architecture [2]. There are few researches on design patterns and architectural languages for microservices [3]. Rademacher et al. point out that the modeling languages for microservices need to be richened in behavior and deployment etc., and should integrate with other related modeling languages [4]. It is necessary to develop new software development methods and software architectural styles for microservice architecture [5].

CBDI-SAE metamodel for SOA 3 has gotten wide attention of academic and industrial circles. Its objective is to provide a detailed concept model together with definitions as a mechanism for clarity and consistency that can form the basis for coherent cross lifecycle asset recording and management [6]. However, it only gives the metamodel consisting of 15 packages and some suggestions with the metamodel, and does not prescribe which models should been built, either not present a modeling process.

In order to solve the problems mentioned above, this paper presents an approach to modeling microservice solutions based on CBDI SAE metamodel for SOA 3.

This paper is structured as follows. Section 2 gives an overview of the microservice application. Section 3 presents a microservice modeling process and prescribes modeling artifacts. Section 4 analyzes related work. Finally, the conclusions are drawn in Sect. 5.

2 Overview of the Microservice Application

The microservice architectural style is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API [7]. Figure 1 illustrates an overview of the microservice applications.

A microservice application includes a microservice application model and a component model for implementing the microservice. A set of microservice forms a microservice application, which runs in the microservice running platform. A microservice is implemented by the components, which run in related component running platforms. Both platforms, which may be Internet environments, cloud platforms, or local area networks, etc. have the ability of real-timely showing their running states and behaviors, and also have the ability of helping the software running on them real-timely to show their internal states and behaviors. In Fig. 1, an evolving microservice application is represented using the microservices circled in a cloud-shaped graph and the relative components.

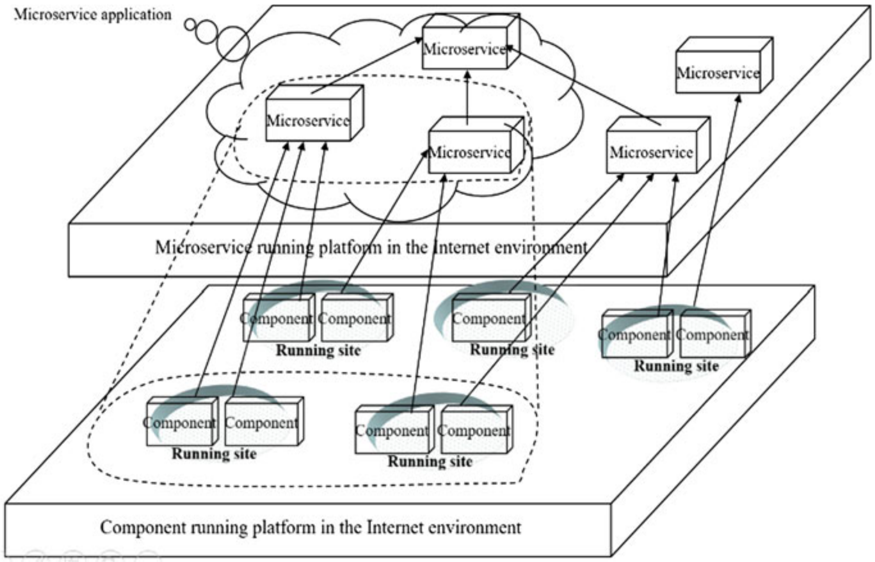


Fig. 1 Overview of the microservice applications

3 Modeling Process

This section discusses how to model microservice solutions according to the ability of CBDI-SAE metamodel for SOA 3. Figure 2 shows framework of the modeling process.

Policy package in [6] contains elements used for describing the policies of an organization. These elements include assertions, assertion variations, constraints, the intended effects of these rules, and the contexts in which these the rules apply. These policies may be applied on various kinds of modeling elements.

Microservice management is used to classify and manage services, is convenient for accessing them, and OASIS UDDI is a specification for managing microservices.

The following part discusses how to build the other models.

3.1 Business Modeling

The goal of business modeling is to create reformed and optimized business models, on the basis of domain analysis, which can agilely adapt to business changes.

Business modeling package in [6] defines business modeling elements, mainly those to which the modeling element Service can be usefully linked. Solution modeling package in [6] can be used to express solution requirements with use cases. Organization package in [6] enables organization units, persons, and posts to be

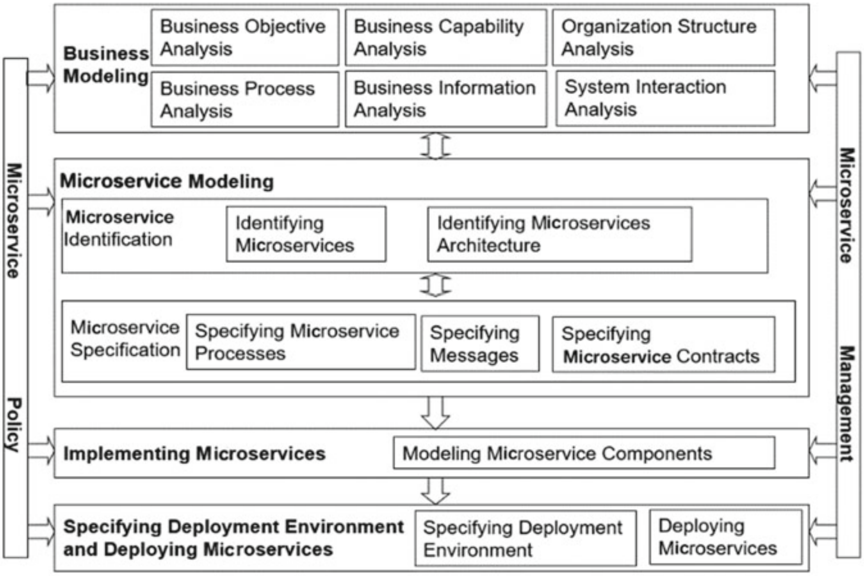

Business Modeling
Business Objective Analysis
Business Capability Analysis
Organization Structure Analysis
Business Process Analysis
Business Information Analysis
System Interaction Analysis
Microservice Modeling
Microservice Identification
Identifying Microservices
Identifying Microservices Architecture
Microservice Specification
Specifying Microservice Processes
Specifying Messages
Specifying Microservice Contracts
Implementing Microservices
Modeling Microservice Components
Specifying Deployment Environment and Deploying Microservices
Specifying Deployment Environment
Deploying Microservices
Microservice Policy
Microservice Management

Fig. 2 Framework of the modeling process

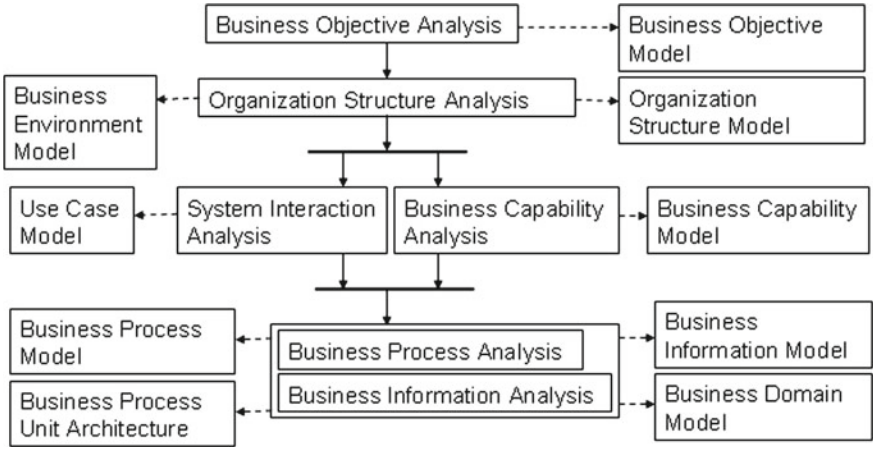

Business Objective Analysis
Business Objective Model
Business Environment Model
Organization Structure Analysis
Organization Structure Model
Use Case Model
System Interaction Analysis
Business Capability Analysis
Business Capability Model
Business Process Model
Business Process Unit Architecture
Business Process Analysis
Business Information Analysis
Business Information Model
Business Domain Model

Fig. 3 Business modeling activities and artifacts

represented in models. Business modelers use the three packages to build business models. Figure 3 shows the business modeling activities and related artifacts.

A business objective is an aim or intended result of a business strategy. Business objective models are the stereotyped UML object models, and objects in business objective models are the instances of Business Objective, Outcome, or Risk in Business Modeling package. These objects are the result of the execution of business services in business processes and provide the analyzers with a mechanism for directly relating business operations back to the objectives of the organization. Business processes, business types and business capabilities in business models must satisfy business objectives.

Business organizations use Organization package to analyze and describe business contexts such as persons and organization units. The modeling artifacts are organization structure models and business environment models, which are depicted with the stereotyped UML class diagram.

System interaction analysis is an activity analyzing the interaction between systems and users of the systems to make certain of the business boundary of microservices for business organizations. The artifacts of the activity are UML use case models, which are built with the modeling elements in Solution Modeling package. Business capability is the power or ability to perform something of value to concerned business. Business capability analysis is an activity making certain of business abilities which systems should have. The artifacts of the activity are business ability models, which are built with modeling elements in Business Modeling package. Business capability models are depicted with the stereotyped UML use case diagrams. Business capability and business interaction may be analyzed simultaneously.

Business processes must satisfy business objectives and business capabilities, and concern the business events triggering them as well as the outcomes and risks caused by them. Analyzers may analyze business processes top-down. The highest level business processes first are identified, with the possible risks and effect on business objectives for executing the business processes. And then analyzers refine the business processes, and identify sub-business processes, which are fulfilled by related roles in organization structure models, and the passed business information between sub-business processes. The business events that trigger (sub-) processes also need to be identified in above analysis. Analyzers may also analyze business processes bottom-up. That is, analyzers first identify basic business processes, and then combine them to realize higher level business processes. Actually, the two kinds of analysis approaches often are applied together. A part of artifacts of the business process analysis is business process models, and another part is business process unit architecture, which are built with modeling elements of Business Modeling package. The former is depicted with the stereotyped UML activity diagram, and the latter with the stereotyped UML class diagram.

When building business process models, analyzers need to identify the follow business units that form business process architecture:

- A process unit designed to provide operations that support one specific business process or sub-process. This unit should be independent from any particular user interface design.
- A capability unit designed to support a particular business capability. This unit should be independent from any particular business process or UI design.

- A core business unit responsible for maintaining records about the instances of particular set of business types. This unit should be independent from any particular business capability, business process, solution, or UI design.
- A utility unit providing common or specialized operations that can be consumed by other business units.

The business information analysis is often needed when building business process models. That is, analyzers need to identify and organize the business information stored or interchanged in business processes. The artifacts of the analysis are business information models, which are built with modeling elements of Type Modeling package. Business information models are depicted with the stereotyped UML class diagrams. According to these models, the microservices to organize and manage the information may be built later, if needed.

A business domain is a major logical partition of an enterprise according to business capabilities and business processes. Microservices may be assigned to a domain as they are identified later. Business domain models are built with modeling elements of Business Modeling package, if needed. Business domain models are depicted with the stereotyped UML class diagram.

3.2 Microservice Modeling

Microservice modeling level separates business modeling level with implementing microservice level in order that built applications can agilely fit the change of both business and implementation technologies.

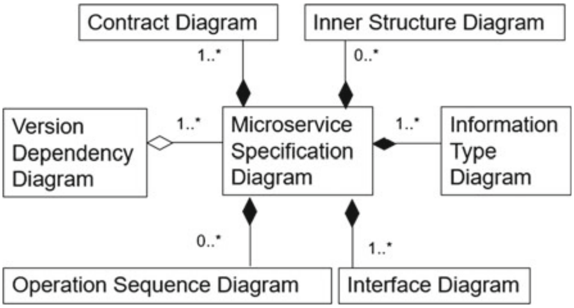
Microservice Identification
The goal of microservice identification is building microservice architecture consisting of microservices and relations between them based on the models in Fig. 3 and existing assets, etc.

The Service package and Software Service in [6] enable some basic information about a microservice to be modeled. Here, modeling elements Service and Software Service in the two packages are all referred to as microservice since they all offer a consumer's capability to a provider through a service interface. In the two packages, a microservice means the idea of the microservice exists, and perhaps a few attributes of the microservice are known, but possibly nothing more.

According to the business models built above and existing assets, analyzers identify microservices convenient for reuse, replacement and development. These microservices are capable of developing applications fast or adapting the change of applications. Thus, the following microservices need to be identified:

- Functional microservices, which are identified according to capability units and core business units.

Fig. 4 Composition of a microservice specification diagram



- Process microservices, which are identified according to business processes. A process microservice consists of functional microservices (as sub-microservices), and manages and controls them.
- Utility microservices, which are identified according to business utility units. They provide common or specialized operations that can be consumed by any other microservice.
- Infrastructure microservices, which provides technical capabilities, rather than business logic. For software microservices, these are the additional capabilities to realize other microservices.

According to business models and identified microservices, the microservice architecture can be built. The microservice architecture is represented with the stereotyped UML class diagram.

The microservice identified above may primarily be specified. Its main information includes:

- Part of basic information, such as a name, a few attributes, roles and function description.
- Primary interfaces and their proposed operations, which are identified according to business events and business information.
- Use case models for describing inside and outside interaction of the microservice. Some operations can be identified from use case models.

Microservice Specification

In order to develop, maintain, use, and access microservices, modelers need to specify microservices according to CBDI SAE metamodel for SOA 3.

The paper gives the composition of a microservice specification diagram that represents a microservice specification, see Fig. 4.

Information type diagrams, which are derived from business information diagrams, describe the information and relations between information used by microservices. An information type diagram is a stereotyped UML class diagram.

A microservice has one or several interfaces. The interfaces are classified as provided interfaces and required interfaces. Modelers need to give all operations of each interface and possible pre- and post-conditions of operations. Operations and their pre- and post-conditions use the information and relations among information above. Interface diagrams represent the information above. An interface diagram is a stereotyped UML component diagram.

An operation often needs to collaborate with other operations. This is expressed as a mandatory operation sequence, which can be thought as schemas for the messages. Operation sequence diagrams represent above information. An operation sequence diagram is a stereotyped UML sequence diagram.

An inner structure diagrams describes the boundary and constitutions of a microservice that has sub-microservices according to a related process microservice. The diagram is a stereotyped UML composition structure diagram.

A microservice contract diagram describes a SLA (Service Level Agreement) that defines the obligations of providers and consumers of a microservice. The SLA must refer to or subsume microservice specifications. A microservice contract diagram is a stereotyped UML class diagram.

A version dependency diagram describes the dependency relations between different versions of a microservice specification. A version dependence diagram is a stereotyped UML class diagram.

Implementing Microservices

It is necessary to model the artifacts implementing microservices. Implementation package and Automation Unit Specification package in [6], whose key concepts are Automation Unit, Technical Interface and Automation Unit Specification, can be used to model and specify the implementation of microservices.

An automation unit describes the implementation of a microservice. On one hand, an automation unit is embodied from a microservice specification according to implementation conditions, and on the other hand, an automation unit is a collection of artifacts which need to be deployed as a group.

The results of modeling implementation of services are represented with the stereotyped UML component diagrams.

3.3 Running Environment and Service Deployment

This section discusses how to model the running environments of implemented microservices and the deployment of the microservices in the running environments. The results of modeling are running environment and deployment models.

A deployment artifact (i.e. a deployable artifact) in Deployment and Runtime package [6] is a physical entity that represents a distinct and deployable part of an automation unit, for example, an executable file or a script. Deployable artifacts are described with UML artifact diagrams.

Technology package in [6] defines modeling elements to model running environments. The modeling artifacts are the diagrams consisting of nodes, execution environments (such as OS), processors and communication protocols etc. which the deployment of automation units relates to.

Deployment and Runtime package in [6] can be used to model a set of objects needed to express microservice deployment and runtime behavior.

The stereotyped UML deployment diagram can represent the running environment and deployment models of microservices.

4 Related Works

OASIS reference model for SOA [8] presents a set of concepts, which are the foundation of building microservices, generally not used to model applications. OMG' SoaML [9] only pays attention to high abstract level modeling, not referring to service implementation and deployment, without a detailed modeling process. CBDI SAE 3 provides a detailed concept model that can form the basis for modeling applications based on services [6].

In the systematic mapping study for microservice architecture, Alshuqayran et al. show that CAMLE may be used to design for service-oriented systems, but it focuses on agent-oriented modeling [5]. Balalaie et al. depict modeling concepts for migration to a cloud-native architecture and a simple migration process for microservice architecture [10]. Vianden and Sun et al. apply UML component diagram, sequence diagram and class diagram to describe microservices [11, 12].

O'Connor et al. give a simple software development process for a microservice architecture with a case study [13]. Levcovitz et al. provide a dependency graph method with simple several steps to identify microservices on monolithic systems [14]. Wei Gu gives a simple metamodel to describe microservice architecture, and the basic ideas on the development process for using the metamodel [15].

5 Conclusions

This paper presents a detailed approach to modeling microservice solutions based on CBDI SAE 3. That approach gives modeling activities and relations between them, definitely points out which models may be built in each activity, and prescribes relations between the models and how to describe the models.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61672046).

References

1. Fowler SJ (2017) Production –ready microservices. O'Reilly Media, Inc.
2. Mazzara M, Khanda K, Mustafin R,Rivera V,Safina L (2016) Microservice science and engineering. In: Proceeding of international conference in software engineering for defense applications, pp 11–20
3. Francesco PD, Malavolta I, Lago P (2017) Research on architecting microservices: trends, focus, and potential for industrial adoption. In: Proceeding of IEEE international conference on software architecture, pp 21–30
4. Rademacher F, Sachweh S, Zündorf A (2017) Differences between model-driven development of service-oriented and microservice architecture. In: Proceeding of IEEE international conference on software architecture workshops, pp 38–45
5. Alshuqayran N, Ali N, Evans R (2016) A systematic mapping study in microservice architecture. In: Proceeding of IEEE, international conference on service-oriented computing and applications. IEEE, pp 44–51
6. Everware-CBDI: CBDI-SAE™ meta model for SOA version 3. http://creativecommons.org/licenses/by/3.0/
7. Fowler M, Lewis J Microservices. thought works. http://martinfowler.com/articles/microservices.html
8. OASIS: reference model for service oriented architecture 1.0. http://docs.oasis-open.org/soa-rm/v1.0/
9. OMG: service oriented architecture modeling language (SoaML) Specification. OMG document: Formal/2012-05-10
10. Balalaie A, Heydarnoori A, Jamshidi P (2016) Microservice architecture enables DevOps: an experience report on migration to a cloud-native architecture. IEEE Softw 33(3):42–52
11. Vianden M, Lichter H, Steffens A (2015) Experience on a microservice-based reference architecture for measurement systems. In: Proceeding of 21st Asia-Pacific Software engineering conference, pp 183–190
12. Sun L, Li Y, Memon RA (2017) An open IoT framework based on the microservice architecture. China Commun 14(2):154–162
13. O'Connor R, Elger P, Clarke PM (2016) Exploring the impact of situational context — a case study of a software development process for a microservice architecture. In: Proceeding of IEEE/ACM international conference on software and system processes, pp 6–10(2016)
14. Levcovitz A, Terra R, Valente MT (2016) Towards a technique for extracting microservices from monolithic enterprise systems. In: Proceeding of brazilian workshop on software visualization, evolution and maintenance, pp 97–104
15. Gu W (2017) Technical practices based on microservice architecture. The engineering report of enterprise architecture innovation institute of PREMET. http://www.primeton.com/read.php?id=2189

A Comprehensive Investigation of BPMN Models Generation from Textual Requirements—Techniques, Tools and Trends

Bilal Maqbool, Farooque Azam, Muhammad Waseem Anwar, Wasi Haider Butt, Jahan Zeb, Iqra Zafar, Aiman Khan Nazir and Zuneera Umair

Abstract Business Process Modeling Languages (BPML's) are continuously getting attraction of software development communities due to the fact of specifying complex business requirements with simplicity. However, the development of business process models from textual requirements through existing BPML's is a time consuming task. In this context, Natural Language Processing (NLP) techniques are commonly applied to automatically generate business process models

B. Maqbool (✉) · F. Azam · M. W. Anwar · W. H. Butt · J. Zeb · I. Zafar · A. K. Nazir
Department of Computer & Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad, Pakistan
e-mail: bilal.maqbool16@ce.ceme.edu.pk

F. Azam
e-mail: farooq@ceme.nust.edu.pk

M. W. Anwar
e-mail: waseemanwar@ceme.nust.edu.pk

W. H. Butt
e-mail: wasi@ceme.nust.edu.pk

J. Zeb
e-mail: jahanzeb@ceme.nust.edu.pk

I. Zafar
e-mail: iqra.zafar16@ce.ceme.edu.pk

A. K. Nazir
e-mail: aiman.nazir16@ce.ceme.edu.pk

Z. Umair
Department of Computer Science, City University of Hong Kong, Hong Kong, China
e-mail: zaziz3-c@my.cityu.edu.hk

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_54

from textual requirements. Business Process Model and Notation (BPMN) is a well-known BPML. This article comprehensively investigates modern techniques, tools and trends for the generation of BPMN models from textual requirements by utilizing NLP techniques. Particularly, a Systematic Literature Review (SLR) is performed to select and evaluate 36 research studies published in the span of 2010–2018. As a result, 11 NLP and 8 BPMN tools are identified. Furthermore, 8 commonly generated BPMN constructs are recognized. Finally, a comparative analysis of NLP and BPMN tools is performed with the help of important evaluation parameters. It is concluded that the existing NLP techniques and tools significantly simplify the process of BPMN models generation from textual requirements. However, the existing approaches are inadequate to be applied in the industries, especially for real-time systems.

Keywords NLP · BPML · BPMN · Natural language processing · Business process model and notation

1 Introduction

Processes in business are essential part of organizations. Business processes provide an overview of all the operations that will be done in the company to provide or create services. By improving the processes in the organization, can increase productivity and efficiency of the products [1].

Business Process Management (BPM) focuses on improving the process to improve productivity, reduce costs, increase service quality, and so on. The company shall identify its resources and processes to implement BPM in their organization [2]. As it is easy for peoples to define business process in natural language since mostly participants are not familiar with designing techniques. Hence, companies have large number of documents written in natural language. Many companies store information in an unstructured way (text documents). Having some automatic tool that can generate process models from natural language documents can make their business more productivity and efficient.

As process modeling is time consuming and need more effort and expertise, that's why many organizations just rely on text documents [1]. Manually developing business processes need participants to conform them to business policies, which is also a difficult task. Solution to these issues is Natural Language Processing, through which we can extract business process from business processes documents and analyze them [3]. Here in 2004 BPMN emerged to control and formalize processes in the companies and is standardized by the Object Management Group (OMG). A better process can increase productivity, reduce costs, increase product quality and employees motivation of any organization. Hence BPMN is used in many organizations as it can improve business process [4]. Business Process Management is the process of modeling, execution, control, automation, optimization and measurement of activities of business flows, to support goals of organization, traversing employees, systems, partners and customers in and beyond the boundaries of organization [4, 5].

According to Dumas BPM have following activities; Process Discovery, Identification, Examination (Analysis), Process Restructuring, Implementation, Monitoring and controlling of business processes [6]. Natural Language is developed naturally without planning and modeling, English language is a well-known example [7]. Due to complexity and ambiguity it is difficult and complex to process natural language. For this many techniques and tools of natural language processing are introduced to perform processing on natural language. NLP is used in the context of BPMN to extract processes form natural language documents and analyze processes [7]. NLP can be helpful to automatically generate business models.

The research [3] is still at an early stage. Lots of further improvements and next level research work needed for authentication of early stage work. The research is at initial stage and just a simple prototype software was build. BPMN models manually made in organizations does not fully compliance with OMG standards [8]. Many BPMN constructs and their sub-varieties are not automatically generated using NLP in previous researches [9, 10], like timer, signal, events have not considered yet for real-time systems. Chunking technique is only being discussed but not used [11]. We have developed following research questions for this SLR:

RQ 1. What are significant researches reported from 2010–2018 where NLP approaches are used to generate BPMN? **RQ 2.** What are the noteworthy NLP tools used & proposed by the researchers for business process discovery during 2010–2018 researches? **RQ 3.** What are the strengths and limitations of NLP approaches and tools have been reported during 2010–2018 researches? **RQ 4.** What are the BPMN modeling constructs auto generated in 2010–2018 researches using NLP approaches? **RQ 5.** What are the noteworthy BPMN modeling tools used & proposed by the researchers for business process discovery during 2010–2018 researches?

2 Development of Review Protocol

2.1 Inclusion and Exclusion Criteria

We define solid criteria to select and reject the researches and define six constraints to guarantee the precision of research question's results. We will select or reject the researches on the ground of following constraints:

(1) Research work would be selected if it is related to our research perspective and supporting research questions findings. Papers will be included that have used natural language processing to auto generate BPMN models. All unrelated researches would be rejected those are not related to our subject and not related to our research questions. (2) Second constraint is to select the research work that is published during 2010 to 2018. We will reject the older work to include the latest research work only to ensure the quality of results extracted. (3) Mostly five famous scientific databases

should include the selected research work i.e. Springer, IEEE, ACM, Elsevier and Taylor & Francis. (4) We will select the researches having vital productive results regarding BPMN models using NLP approach. All the research work having no noteworthy values would be rejected. (5) All the nominated researches must have some defined results and must be sustained by concrete evidences. All the researches having unverified or unauthentic research method would be rejected. (6) We will discard all the studies that are matching to each other for the specific research situation and single one of them would be selected to avoid overfitting of results.

2.2 Search Process Details

Five scientific databases, Springer, ACM, IEEE, Taylor & Francis and Elsevier containing high impact factor conference proceedings and journals are selected and different search terms like BPMN, NLP and Natural Language Processing etc. were used for this SLR. The results of the selected search terms are shown in the Table 1. Moreover, Table 1 also summarize the search terms results with year filter.

"2010–2018" filter was used for the all search terms to filter the research work published during 2010–2018. Moreover, AND/OR operator was used for the search terms containing more than one word to get the more search results but AND/OR operator don't tell about the quality of search results. AND/OR operator has no scope for a single word search term so "0" is used for single word search term.

Table 1 Summarized search terms with year

#	Search term	Operator	No. of search results (2010–2018)				
			IEEE	Springer	Elsevier	ACM	Taylor & Francis
1	Business process model and notation	AND	85	135	203	26	37
		OR	334	3,072	6,251	12,438	19,833
2	Natural language processing	AND	10,666	3,271	3,959	5,810	967
		OR	12,475	15,081	20,973	9,985	12,469
3	Extracting business processes using NLP	AND	0	0	0	0	0
		OR	9	302	515	976	217
4	NLP tools and techniques	AND	47	222	211	21	50
		OR	257	1,190	1,889	3,652	1,187
5	BPMN modeling tools	AND	0	9	5	0	0
		OR	94	1,165	931	8,701	141

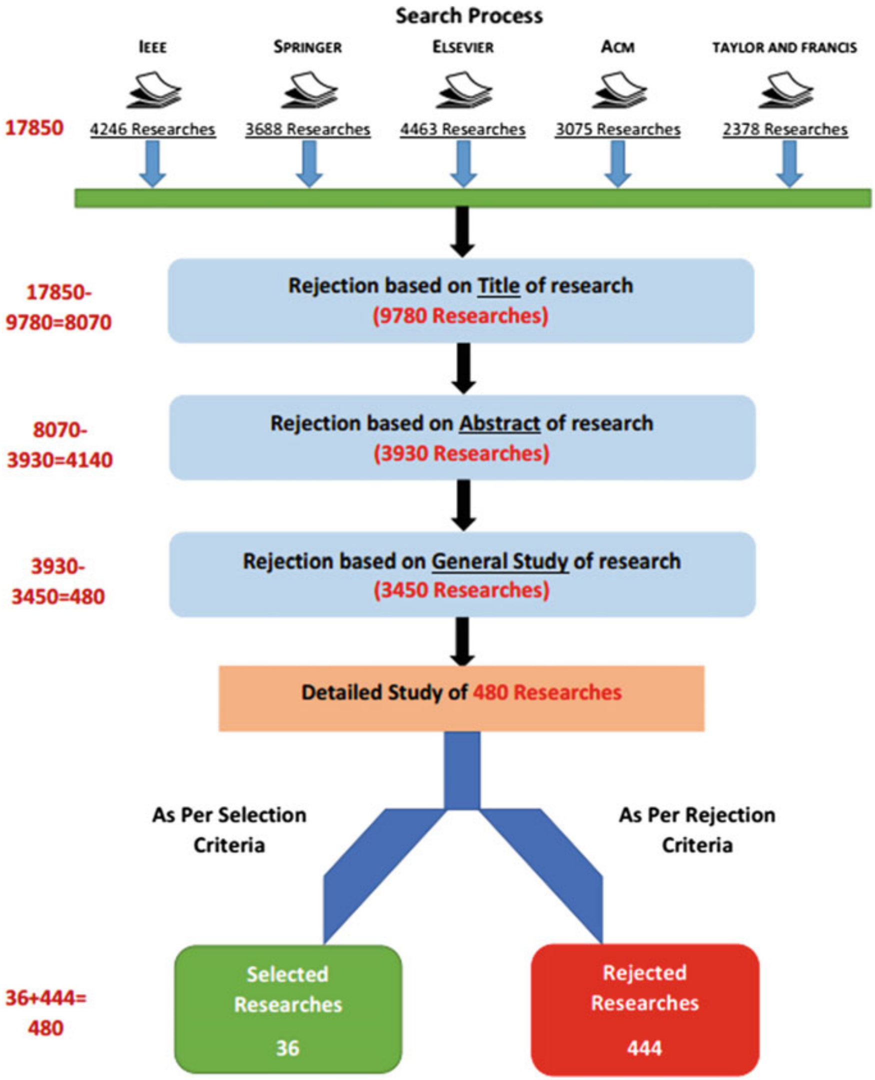


Fig. 1 Search process

Further advance search is also used to further refine the results of our search keywords from the five selected scientific databases. Following steps were performed to search the scientific databases and detail is also shown in the Fig. 1.

(1) We identify several "search terms' in five scientific databases, scrutinize them and selected almost 17850 search results according to the rejection and selection criteria. (2) 9780 studies are discarded by evaluating their Title according to the rejection and selection criteria. (3) 3930 studies are discarded by evaluating their

Table 2 Databases versus journal and conference papers

#	Scientific database	Category	Research works	Number of references
1	IEEE	Journal	[12]	1
		Conference	[3, 9, 13–18]	8
2	Springer	Journal	[1, 4, 5, 7, 8, 12, 19]	7
		Conference	[10, 20–25]	7
3	Elsevier	Journal	[2, 26–28]	4
		Conference	–	–
4	ACM	Journal	–	–
		Conference	[29–31]	3
5	Taylor & Francis	Journal	–	–
		Conference	–	–
6	Others	Journal	[11, 32–35]	5
		Conference	[36]	1

Abstract according to the rejection and selection criteria. (4) 4140 studies are selected to perform general study and different sections are studied. Further, 3450 research works are rejected according to the rejection and selection criteria based on general study. (5) We accomplish thorough study of 480 studies and 444 researches are further rejected. (6) At the last, 36 researches were selected having completely agreement with our rejection and selection criteria.

2.3 Quality Checking

Quality cehcking principles have been established to recognize the significant results from the nominated researches. The established principles also describe the trustworthiness of individually nominated researches and their critical results:

(1) The facts evaluation of study is grounded on the tangible evidences and theoretic understanding without some unclear declarations. (2) The proof of study has been achieved using appropriate justification approaches e.g. case study etc. (3) The study delivers evidence around the NLP approches used to perform BPMN models generaton. (4) Most recent studies i.e. 56.76% from 2015 to 2018 and overall 72.97% from 2013 to 2018 are included to examine the most recent business proesses and NLP tools approches. (5) Five quality scientific databases, Springer, ACM, IEEE, Taylor & Francis and Elsevier were selected to improve the accuracy and qulaity of our search results (Table 2).

Table 3 Categorization of researches

#	Category	No. of researches	Research identification
1	General	6	[8, 15, 17–19, 24]
2	Health care	10	[2, 13, 16, 22, 26, 27, 29, 33–35]
3	Information technology	16	[1, 3–5, 7, 9–12, 14, 20, 21, 25, 31, 32, 36, 37]
4	Supply chain system	4	[23, 28, 30, 36]

2.4 Data Extraction Procedure for Selected Research Works

Definition of Categories
Four categories are formulated to consolidate the chosen researches, shown in (Table 3). This organization drastically improves the correctness of the answers of our studies questions. Health care, Information technology and supply chain category includes papers who have applications in medical, software industry and transportation field, respectively. Papers those have not clearly application in any defined field are kept in general category.

Data extraction and synthesis
Data synthesis after data extraction is used to get the responses of our research questions from the selected particular researches as shown in Table 4.

3 Results

3.1 NLP Tools and Techniques

Different NLP approaches have been established which benefits in processing natural language. We reviewed the researches to find out the most studied and used approaches. Results given in the Table 5 shows the popularity of each approach.

Paper [7, 9] have used both tokenization and POS tagging techniques, paper [22, 29] have used tokenization, POS tagging and parsing techniques in their BPMN models generation. Chunking technique is only discussed in one paper [11]. Tokenization is the procedure of dividing the given textual requirements into units/blocks termed as tokens [10]. Parsing is the procedure of allocating structural descriptions to classifications of the words in natural language text [10]. Chunking is the process of categorizing ideas in a hierarchy [11]. Part-of-speech (POS) is the process of marking up a word to its corresponding part of speech [10].

We have identified NLP tools from the selected researched in Table 6. Identified tools will help the practitioners in their natural language processing. All the tools are

Table 4 Data extraction and data synthesis

#	Description	Details
1	Bibliographic data	Author, title, publisher details, publication year and kind of research (i.e. conference or journal publication)
Extraction of data		
2	Outcomes, assumptions	Results attained from the particular study. If any assumption exists for the confirmation of outcomes
3	Summary, information collecting	The elementary idea and foundation of particular study, qualitative or quantitative
Synthesis of data		
4	Grouping	Belongs to only single and already defined categories in Table 3
5	Identification of approaches	NLP approaches used by researchers to perform the natural language processing of business process documents (Table 5)
6	Consideration of tools	Tools used or proposed by researchers to perform the natural language processing of business process documents (Tables 6 and 8)
7	Identification of BPMN constructs	Related to BPMN construct in Table 7

Table 5 NLP techniques

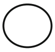
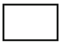
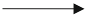
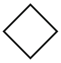
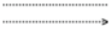
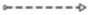
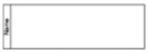
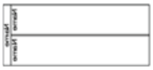
Sr. No.	Technique	Purpose	References	Total
01	Tokenization	Split natural language sentence into tokens	[7, 10, 11, 13, 16, 29–32]	8
02	Parsing	For structural descriptions assignment to words sequence of in NL	[7, 9, 10, 13, 14, 20, 21, 29, 31, 33, 35]	11
03	Chunking	For the hierarchy of ideas	[11, 15]	2
04	Part-of-speech (POS)	Marking up a word to its corresponding part of speech	[7, 10, 16, 20–22, 29, 31, 37]	9

Table 6 Identified NLP tools

Sr. No.	Tools	Purpose	References
01	GATE	For solving text processing problem	[11]
02	CRF++	For segmenting/labeling sequential data, text chunking etc.	[11]
03	Ling-pipe	Tool kit for processing text using computational linguistics	[11]
04	Stanford POS tagger	Used for POS tags to each word	[11, 20, 31]
05	Stanford parser	Used for syntax parsing	[3, 7, 9–11, 14, 21, 22, 29, 31, 32]
06	FrameNet	For semantic analysis	[3, 10, 11]
07	WordNet	It is a lexical (word) catalogue language. It clusters words into sets of substitutes (synonyms) known as synsets	[3, 9–12, 15, 22, 31, 32, 37]
08	CoreNLP	It offers a technology tools for the set of human language. A comprehensive variety of linguistic analysis tool with integrated toolkit of NLP	[3, 22]
09	NLTK tagger	This tools is used to produce syntax tree	[9, 32]
10	Bag-of-words	Used in methods of document classification where the (frequency of) occurrence of each is to be used	[10, 34, 37]
11	VerbNet	It is categorized independent of domain, wide-coverage lexicon of verb with mapping facility to other resources, e.g. WordNet	[9–11, 32]

described in terms of NLP approaches used by the authors. Stanford parser is used my many researches, can be seen in the Table 6.

Table 7 BPMN modelling constructs

Sr. #	Modelling constructs	Notation	No. of researches	Research identification
1	Event		5	[9, 10, 12, 26, 31]
2	Activity		8	[3, 7, 9, 10, 12, 13, 26, 31]
3	Sequence flow		7	[3, 9, 10, 12, 13, 26, 31]
4	Gateway		7	[3, 9, 10, 12, 13, 26, 31]
5	Association		3	[12, 26, 31]
6	Message flow		3	[10, 26, 31]
7	Pool		4	[10, 12, 13, 31]
8	Lane		2	[10, 31]

3.2 BPMN Model Constructs

Table 7 shows a Business Process Constructs list that can be represented by business process modelling notation (BPMN) that is modelled using NLP. An "**Event**" is what happens due to some act of process, it can effect flow of model and always have a trigger (cause) or result (impact) that initiate it to do some operation. Events are donated with open center circle. Many verities of these constructs are not yet generated from natural language requirements, like complex gate, event based gate, timer event, conditional event, signal event and many more are not generated.

"**Activity**" is a term used to depict some work that an organization do in a process. Activities can be divided into sub-parts: Task and Sub-Process. Activity is represented by rounded rectangle. Activities can be used in both Choreographies and in standard Processes. A "**Sequence Flow**" show the order in which activities will perform operation in process. "**Gateway**" controls the separation and conjunction of Sequence Flows in processes. It can control branching, merging, forking, and joining, of paths. "**Association**" is used to connect BPMN constructs and their information flow. "**Message Flow**" shows flow of messages between participants. "**Pool**" is use to depict participant in a teamwork. "**Swim lane**" graphical represent a container for participants doing activities from other Pools. "**Lane**" represents a sub-partition within a Process. Lanes can categorize and organize different activities.

Table 8 BPMN modelling tools

Sr. #	Tools	No. of researches	Research identification
1	WebRatio	2	[24, 28]
2	BizAgi process modeler	6	[2, 10, 18, 21, 23, 28]
3	BPMS	5	[11, 17, 28, 30, 36]
4	Activiti BPM	3	[17, 18, 28, 30]
5	Oracle BPMN	2	[10, 36]
6.	JESS	2	[17, 26]
7.	CLIPS	1	[26]
8.	JBoss rules	2	[26, 28]

3.3 BPMN Modelling Tools

Table 8 lists the research papers which mentioned tools for BNMN modelling.

"**Web Ratio**" is used to graphically depict enterprise Web application and mobile App development production and used for modelling Business Processes. "**BizAgi Process Modeler**" is used for modelling Business Processes. **Business Process Management Systems** (BPMS) provide support for the business process (BPs) lifecycle, from modelling to executing and evaluating BPs. "**Activiti BPM**" is battle-tested Business Process Management (BPM). Organizations across the world depend on the open source platform. **CLIPS, JESS**, and **JBoss**, provide platforms for rule based execution, control and management.

4 Comparison

In this section, we have evaluated and analyzed NLP tools identified against many factor like complexity of tool, pricing (open source or licensed), and etc. described in selected research studies. In Table 9, we extracted those tools with are proposed or used by the researchers. These tools are already listed in Table 6. Complexity means whether is handles complex natural language inputs or not. Open source means that the software is free of cost or not. Compatibility mean is work done in this software is compatible with other NLP tools or its previous/new versions. Among other properties, tools can be implemented in different programming languages; Programming language of tools are also identified in Table 9.

In this section, we have evaluated and analyzed BPMN modeling tools identified against many factor like complexity of tool, pricing (open source or licensed), etc. and their particular outcomes (output format) described in selected research studies. In Table 9, we extracted those tools with are proposed or used by the researchers. These tools are already listed in Table 8. Complexity means tool is easy to use or not. Open source means that the software is free of cost or not. Compatibility

Table 9 Comparison of identified NLP tools

Sr. No.	Tools	Complexity	Open source	Compatibility	Programming language	References
01	GATE	✓	✓	✓	Java	[11]
02	CRF++	×	✓	×	Python	[11]
03	Ling-pipe	×	×	✓	Java	[11]
04	Stanford POS tagger	✓	✓	✓	Java	[11, 20]
05	Stanford parser	✓	✓	✓	Java	[3, 7, 9–11, 14, 21, 22, 29, 32]
06	FrameNet	✓	×	✓	Python	[3, 10, 11]
07	WordNet	✓	×	✓	Python	[3, 9–12, 15, 22, 32, 37]
08	CoreNLP	✓	✓	✓	Java	[3]
09	NLTK tagger	✓	✓	✓	Python	[9, 32]
10	Bag-of-words	✓	✓	✓	Python	[10, 34, 37]
11	VerbNet	✓	✓	✓	Python	[9–11, 32]

Table 10 Comparison of identified BPMN modeling tools

Sr. No.	Tools	Complexity	Open source	Compatibility	Modeling	Validation	Platform	Reference
1	WebRatio	×	×	×	✓	✓	Windows/Linux/Mac	[24, 28]
2	BizAgi process modeler	×	×	✓	✓	✓	Windows/Linux/Mac	[2, 10, 18, 21, 23, 28]
3	BPMS	✓	×	×	×	×	Web Portal	[11, 17, 28, 30, 36]
4	Activiti BPM	×	✓	×	✓	✓	Cross-platform	[17, 28, 30]
5	Oracle BPMN	×	×	✓	✓	✓	Windows/Linux/Mac	[10, 36]
6	JESS	✓	×	×	×	×	Plugin	[17, 26]
7	CLIPS	✓	×	×	×	×	Plugin	[26]
8	JBoss rules	✓	✓	✓	×	×	Windows	[26, 28]

mean is work done in this software is compatible with other BPMN modeling tools or its previous/new versions. Among other properties, tools can be implemented in different programming languages; Modeling mean whether it supports graphical modeling of BPMN model or not. Output mean what format or kind of result can be achieved from using this tool. × mean that factor is not present and ✓ mean that factor is present in corresponding tool (Table 10).

In this section, we have evaluated and analyzed techniques tools and their particular outcomes after apply those techniques on natural language requirements described in selected research studies. For chunking techniques GATE tool is mostly used for segmenting/labeling sequential data, text chunking etc. and text chunks are received as an output of this process. In parsing technique, Stanford POS Tagger structural descriptions assignment to words sequence of in NL and at the end we get structural division of words. For tokenization purpose, WordNet is used in many papers for split natural language sentence into tokens. Marking up a word to its corresponding part of speech, POS technique is used and for achieving this purpose successfully CoreNLP tool is used in many research work that are reviewed in this document.

5 Discussion and Limitations

36 important researches from 2010–2018 have been identified. 7 important NLP tools used and proposed by researchers from the selected studies are shown in the Sect. 3.1. Table 6 further describes each tool with its identifying researches, website, occurrence count in the selected research studies and their purposes. Only 8 BPMN constructs are auto generated using NLP. 8 important BPMN modeling tools used and proposed by researchers from the selected studies are shown in the Sect. 3.4. Table 8 further describes each tool with its identifying researches and the occurrence count in the selected research studies.

Many BPMN constructs and their sub-varieties are not automatically generated using NLP in previous researches. Chunking technique is only being discussed but not used. BPMN is dynamic modeling language and mostly used in the industry. Although proper tools and techniques to perform any SLR is used but still, there are minor limitations as discussed below:

(1) Even though we used proper search terms, inclusion-exclusion criteria to select one of the five famous scientific databases should include the selected research work i.e. Springer, IEEE, ACM, Elsevier and Taylor & Francis but still there is a minor chance that some fraction of researches in other databases are missed. (2) Some research studies have different title and content so there is a chance to miss the researches based on title based rejection. Although we used the advanced search to check the abstract and other parts of research too but still there is a minor chance that some fraction of researches is missed.

6 Conclusion and Future Work

This article investigates the latest techniques, tools and trends for the generation of Business Process Model and Notation (BPMN) models from textual requirements by utilizing Natural Language Processing (NLP) techniques. Particularly, 36 studies published in the span of 2010–2018 are selected with the help Systematic Literature Review (SLR) guidelines. The identified research studies are comprehensively evaluated to present 11 NLP and 8 BPMN tools. Moreover, 8 commonly generated BPMN constructs are recognized. Furthermore, a comparative analysis of NLP and

BPMN tools is performed to highlight the major strengths and weaknesses. Finally, it is concluded that the existing NLP techniques and tools provide strong foundation for the generation of BPMN models from textual requirements. However, the current approaches are inadequate to be applied in the industries especially for real time systems. Therefore, it is required to develop sophisticated approaches and tools in view of the complex requirements of real time systems. In future, we intend to extend this research by considering other well-known Business Process Modeling Languages (BPML's) like Event-driven Process Chain (EPC) etc.

References

1. Bocciarelli P, D'Ambrogio A (2014) A model-driven method for enacting the design-time QoS analysis of business processes. Softw Syst Model 13(2):573–598
2. Corradini F et al (2017) A Guidelines framework for understandable BPMN models. Data & Knowledge Engineering
3. Sintoris K, Vergidis K (2017) Extracting business process models using natural language processing (NLP) techniques. In: 2017 IEEE 19th conference on business informatics (CBI), vol 1. IEEE
4. Grolinger K et al (2014) Integration of business process modeling and Web services: a survey. Serv Oriented Comput Appl 8.2(2014):105–128
5. Branco MC et al (2014) A case study on consistency management of business and IT process models in banking. Softw Syst Model 13.3(2014):913–940
6. Dumas M et al (2013) Fundamentals of business process management. Springer Publishing Company, Incorporated
7. Leopold H (2013) Natural Language in Business Process Models: Theoretical foundations, techniques, and applications. Lecture Notes in Business Information Processing, vol 168. Springer International Publishing
8. Hashmi M et al (2018) Are we done with business process compliance: state of the art and challenges ahead. Knowl Inf Syst
9. Epure EV et al (2015) Automatic process model discovery from textual methodologies. In: 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS). IEEE
10. Friedrich F, Mendling J, Puhlmann F (2011) Process model generation from natural language text. In: International conference on advanced information systems engineering. Springer, Berlin, Heidelberg
11. Osman C-C, Zalhan P-G (2016) From natural language text to visual models: a survey of issues and approaches. Inf Econ 20(4):44
12. Pittke F, Leopold H, Mendling J (2015) Automatic detection and resolution of lexical ambiguity in process models. IEEE Trans Softw Eng 41(6):526–544
13. de AR Gonçalves JC, Santoro FM, Baião FA (2010) A case study on design-ing business processes based on collabora-tive and mining approaches. In: IEEE of 14th international conference on sup-ported cooperative work in design (CSCWD 2010), pp 611–616
14. Kumar N, Singh M (2014) Stanford parser based approach for extraction of Link- Context from non-descriptive anchor-text. In: Proceedings of 3rd international conference on reliability, infocom technologies and optimization, Noida, pp 1–6
15. Pirapuraj P, Perera I (2017) Analyzing source code identifiers for code reuse using NLP techniques and WordNet. In: 2017 Moratuwa engineering research conference (MERCon), Moratuwa, pp 105–110

16. Motahari-Nezhad HR et al (2016) RFPCog: linguistic-based identification and mapping of service requirements in request for proposals (RFPs) to IT service solutions. In: 2016 49th Hawaii international conference on system sciences (HICSS). IEEE
17. de Moura JL et al (2017) Test case generation from BPMN models for automated testing of web-based BPM applications. In: 2017 17th international conference on computational science and its applications (ICCSA). IEEE
18. Medoh C, Telukdarie A (2017) Business process modelling tool selection: a review. In: 2017 IEEE international conference on industrial engineering and engineering management (IEEM), Singapore, pp 524–528
19. van der Aalst WMP, la Rosa M, Santoro FM (2016) Business process management: Don´t forget to improve the process! Business & Information Systems Engineering 58(1):1–6
20. Manning CD (2011) Part-of-Speech tagging from 97% to 100%: is it time for some linguistics? In: CICLing 2011, Tokyo, Japan, 20–26 Feb 2011, pp 171—189
21. Mishra A, Sureka A (2015) A graph processing based approach for automatic detection of semantic inconsistency between BPMN process model and SBVR rules. In: International conference on mining intelligence and knowledge exploration. Springer, Cham
22. Van der Aa H, Leopold H, Reijers HA (2015) Detecting inconsistencies between process models and textual descriptions. In: International conference on business process management. Springer, Cham
23. Yongchareon S et al (2010) BPMN process views construction. Database systems for advanced applications. Springer Berlin/Heidelberg
24. Brambilla M, Butti S, Fraternali P (2010) Webratio bpm: a tool for designing and deploying business processes on the web. Web Eng:415–429
25. Ferreira RCB et al (2017) Assisting process modeling by identifying business process elements in natural language texts. In: International conference on conceptual modeling. Springer, Cham
26. Fan S et al (2016) A process ontology based approach to easing semantic ambiguity in business process modeling. Data Knowl Eng 102 (2016):57–77
27. Karatzoglou A, Feinerer I (2010) Kernel-based machine learning for fast text mining in R. Comput Stat Data Anal 54(2):290–297
28. Delgado A, Calegari D, Arrigoni A (2016) Towards a generic BPMS user portal definition for the execution of business processes. Electron Notes Theoret Comput Sci 329:39–59
29. Elstermann M, Heuser T (2016) Automatic tool support possibilities for the text-based S-BPM process modelling methodology. In: Proceedings of the 8th international conference on subject-oriented business process management. ACM
30. Cheikhrouhou S et al (2013) Toward a time-centric modeling of business processes in BPMN 2.0. In: Proceedings of international conference on information integration and web-based applications & services. ACM
31. Ferreira RCB, Thom LH, Fantinato M (2017) A semi-automatic approach to identify business process elements in natural language texts. In: Proceedings of the 19th international conference on enterprise information systems
32. Riefer M, Ternis SF, Thaler T (2016) Mining process models from natural language text: A state-of-the-art analysis. Multikonferenz Wirtschaftsinformatik (MKWI-16), March:9–11
33. Shaalan K (2010) Rule-based approach in Arabic natural language processing. Int J Inf Commun Technol (IJICT) 3(3):11–19
34. Crowston K, Liu X, Allen EE (2010) Machine learning and rule-based auto-mated coding of qualitative data. In: Proceedings of the American Society for Information Science and Technology, vol 47, no 1
35. Hogenboom F, Frasincar F, Kay-mak U (2010) An overview of approaches to extract information from natural language corpora. In: 10th Dutch-Belgian information retrieval workshop (DIR 2010), pp 69–70
36. Caporale T (2016) A tool for natural language oriented business process modeling. ZEUS
37. Li J, HJ Wang, Bai X (2015) An intelligent approach to data extraction and task identification for process mining. Inf Syst Front 17.6:1195–1208

A Culture-Based Profile Model of Software Evaluators

Khaled Hamdan, Boumediene Belkhouche and Peter Smith

Abstract The paper proposes a new approach to software evaluation, which takes into consideration cultural factors. We use Profile Theory to develop a model that captures the essential technical and cultural characteristics of contextually effective software evaluators. These evaluator-defining characteristics include cultural, organizational, technical, and individual attributes, and relationships among them. We used surveys and literature to identify the prevalent characteristics and then defined a formal model. An illustrative example is elaborated to show how our model can be integrated in a CASE tool. We surmise that identifying the profile of software evaluators is a necessary step to ensure the effectiveness and validity of the evaluation of software systems.

Keywords Cultural attributes · Profile theory · Team diversity · Software project performance · Software evaluation

1 Introduction

Previous research has shown that cultural diversity has a significant impact on software development [1–3]. There is little doubt that software evaluation is a critical step in the life cycle of any system, yet, current methods of software evaluation do not take into account the cultural diversity of the software development team. To produce a reliable evaluation, we need evaluators who are not only technically-competent, but also are keenly aware of the cultural context as defined by its various dimensions [4].

K. Hamdan · B. Belkhouche (✉)
United Arab Emirates University, Al Ain, UAE
e-mail: b.belkhouche@uaeu.ac.ae

K. Hamdan
e-mail: Khamdan@uaeu.ac.ae

P. Smith
Sunderland University, Sunderland, UK
e-mail: peter.smith@sunderland.ac.uk

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_55

Cultural attributes and their differences affect software evaluation among other software development activities [6]. Culture entails elements that come together to form the person or the society. It comprises factors such as knowledge, beliefs, values, traits, experiences, language and religion that make up a community, lifestyle and its way of thinking [7]. Culture can be considered as a set of denotative elements (e.g. beliefs), connotative elements (e.g. attitudes, norms and values), and practical knowledge shared by a group of people who participate in a common social structure [3]. Cultural constraints may set boundaries on software evaluation approaches. They may result in several implications, such as misunderstandings, and inefficiencies in the software evaluation approaches and processes. To be effective, a software evaluation approach should not be considered as an independent or an explanatory approach, but rather as providing a supportable relation and fit with relevant cultural aspects among other organizational features. It is important to identify the impact of the cultural context dimensions on software evaluation and consider associations between the impacts of some cultural dimensions on the software evaluation [8].

Evaluation, as a goal-oriented process, is dependent on the current knowledge of science and the methodological standards [9]. A software evaluation is a type of assessment that is used to determine how well software programs fit for the needs of a given client. An evaluator should bear in mind factors affecting evaluation when developing software. First, the user's cultural background and his/her contextual appraisals are seen as crucial. A cultural disconnect between the user and the evaluator will make it difficult for the evaluator to test an appropriate software for the user. Therefore, the tested software design/production must match the users' target culture. The second factor is the evaluator's competence, which is based on his knowledge and experience in the software design. Finally, the evaluator should take into account whether the product is destined for international or local users to avoid cultural misunderstandings and subsequent problems [4, 5]. That is, an evaluation methodology must take into consideration various human factors, e.g., people's knowledge and skills, learning, and performance capabilities, and compatibilities in diverse software development environments. A software product is a "mirror" of knowledge/skills of software developers, i.e., "a state of the art" of their education/intelligence. People cannot produce more than they know and/or could know [10].

To address issues related to culture and its impact on software evaluation, we develop a culture-based model of software evaluators using Profile Theory. Our model can then be integrated in a Computer Aided Software Engineering (CASE) environment to support and enhance the evaluation process. The model incorporates characteristics of an evaluator involving culture, programming languages, experience in the field, and values that make up a community's way of thinking. We believe that cultural background of the evaluator and the user are necessary to have a global knowledge of the system being developed. Based on our study that surveyed a number of software and technical firms, we identify five characteristics of evaluators that may have an impact on software project cost execution and can differentiate the performance of suitable evaluators.

2 Software Quality and Testing

Quality is typically defined in terms of conformance to specification and fitness for purpose. In his book, *Quality is Free* (1980), Philip Crosby states that: “do things right in the first place, and you won’t have to pay to fix them or do them over”. Crosby formed the principle: “do it right the first time” (DIRFT). His belief was that an organization that has a good quality management system will realize benefit more than pays for the cost of the quality system and thus “quality is free” [11]. That is, it is less expensive to do it right the first time than to pay for rework and repairs. A number of approaches to software quality and testing have been proposed [12]. For example, Fitzpatrick, Smith and O’ Shea proposed an approach which they called the “Software Quality Star” [13, 14] which considers a number of factors relating to the World Wide Web, including visibility, intelligibility, credibility, engageability, differentiation, proprietor development and maintenance. The measurement of the domain of evaluation software systems is complex. The lack of a uniform measurement protocol exacerbates this complexity [15].

Evaluation is one of the key factors required for planning and control during the whole project lifecycle. In order to plan and control software products, we need to increase insight by improving quality, which would enhance control, thereby leading to on-time projects. You can use different techniques to evaluate the quality of a software product. Conformance can be defined as a measure of appropriateness of a system (e.g. performance, design, and behavior) to a specified standard. Here, it is proposed that the use of profile theory to help in obtaining a common understanding in the protocol of evaluation.

Two typical approaches to software evaluation are the criteria-based method driven approach and the situated method. The criteria-based approach is a quantitative assessment of the software in terms of sustainability, maintainability and usability. This can inform high-level decisions on specific component for software improvement [16, 17]. A method driven approach emphasizes functionality and a range of data quality. The situated approach supports empowerment software design, construction, whereas methods are produced to match specific organizational situations.

3 Evaluation-Based Model

It is essential to define common characteristics, which contribute to software evaluation. Numerous studies showed the impact of non-technical aspects of evaluators on software projects [18, 19]. Because of workforce diversity in organizations, evaluators’ culture plays a critical role in work performance. Their characteristics and their interrelationships must be explicitly addressed for organizations to perform their tasks smoothly and successfully. An organization will be more productive when its values are shared by its teams. Furthermore, organizational culture incorporates a set of assumptions, beliefs, and values which guide the function of members of the organization [5].

Common factors that influence the software evaluation process are evaluators' culture, their characteristics, their organizational types, and specialized skills.

Combining evaluators' characteristics with other variables is intricate, involving quantitative measures of capability. The completeness property is significant to identify the primary profile factors and it is equally important for these to be included into the profile description. One major issue that still needs to be addressed is how to use profile theory to describe and measure software evaluation.

The evaluation factors tree was then chosen to indicate the weight of each factor and was also used for the notation of profile description. The tree, in its entirety, is important for the identification of the essential profile factors to be incorporated into the profile as a description. It is then possible to use this notation presented in a profile theory to describe the evaluation process. For example, factors that represent evaluations $f(e)$ could be expressed as follows:

$$f(e) = \{< \varepsilon_1, C, \omega_1) >, < \varepsilon_2, S, \omega_2 >, < \varepsilon_3, O, \omega_3 >, < \varepsilon_4, Q, \omega_4 >\}$$

where:

C Evaluator's culture such as intercultural intelligence, cultural behavior- bias or impartial, human languages application domain.

S Evaluator's system expertise & skills levels, Technical, managerial aspects and system complexity. Software quality characteristics can be completeness, reliability, conformance and intelligibility.

O Organizational such as competencies and authorities, credibility, soundness and support.

Q Evaluator characteristics are like soft skills, decision maker, capabilities, compatibilities in software development and knowledge.

ε_i Factor existence, $\varepsilon = 1$for existence, $\varepsilon = 0$ nonexistence;

ω_i Show the weight of sub-factor(s)—weight is divided equally into i, based on importance or priority; ω1, ω2, ω3, and ω4, are the weights for evaluator characteristics, skill levels, organizational, and culture, respectively (Fig. 1).

These assumptions typically include personality traits, power relationships, behaviors and values. When quality characteristics are integrated for capability and compatibility's sake, they provide a definition of how important evaluator capability needs to be in order to fit the product.

An aspect profile capability (V_i) for knowledge/skills conformance is defined as a sum of all factor capabilities:

$$V_i = \sum_{j=1}^{m_i} \omega_j \left(\frac{\varepsilon_j}{\varepsilon_j^{(())}} \right) \left(\frac{v_j}{v_j^{(())}} \right)^2 \tag{1}$$

where for the jth factor:

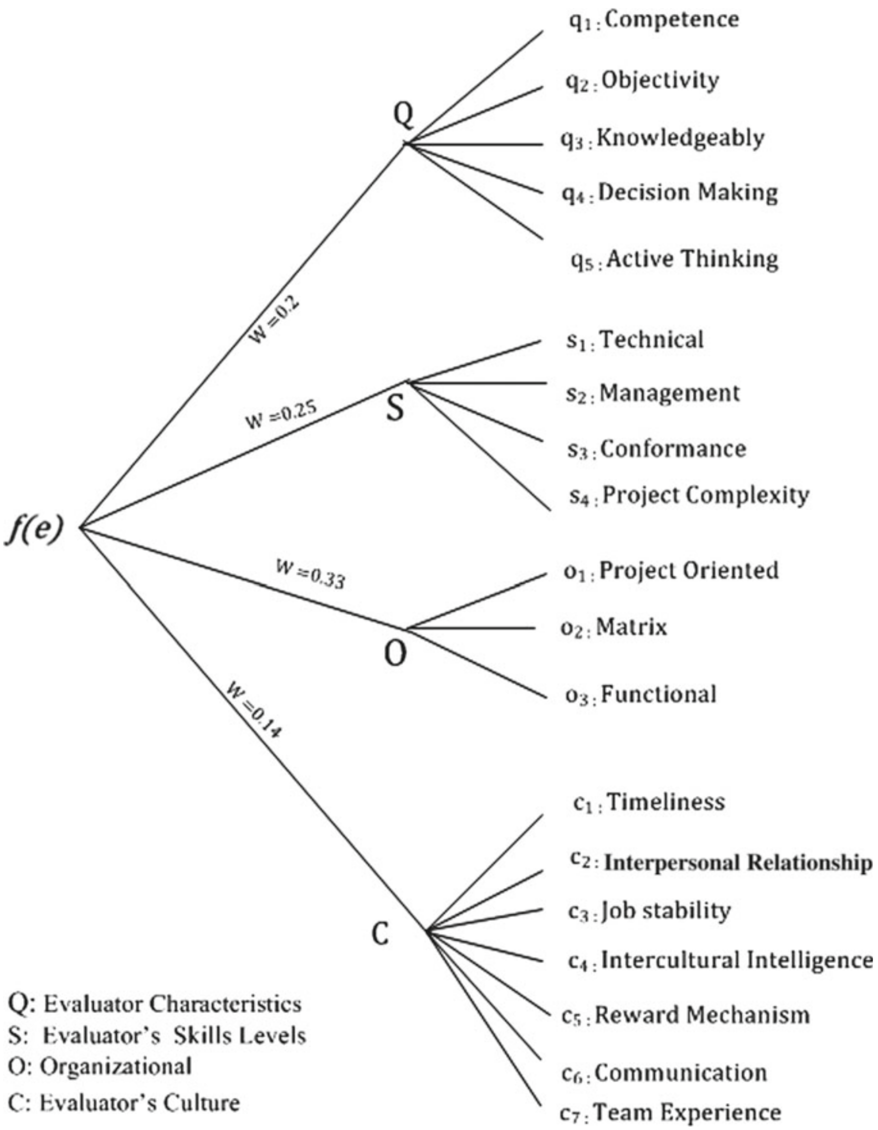


Fig. 1 Evaluator's relationship model

ε_j factor existence such as $\varepsilon = 1$, non-existence $\varepsilon = 0$
$\varepsilon_j^{(())}$ required experience, $\varepsilon_j^{(())} \neq 0$
v_j existing level of factors
$v_j^{(())}$ required level of factors, $v_j^{(())} \neq 0$
ω_j the jth factor (preference/importance) weight

m_i number of factors that are used for the description of evaluation aspects

In order to determine compatibility of the available knowledge/skills conformance with the required one we use the following compatibility measure [20, 21]:

$$W_i = \prod_{j=1}^{m_j} \left(\frac{\varepsilon_j}{\varepsilon_j^{(())}}\right)\left(\frac{v_j}{v_j^{(())}}\right)^2 \tag{2}$$

where W_i is the evaluation aspect compatibility.

4 Survey Components

A survey was carried out to address which factors are perceived as characterizing evaluators and cultural characteristics involved in the evaluation process. We measured Evaluator's characteristics through five items and we assessed each one using four sub-items, while we measured cultural characteristics using seven items and for each item, we used four sub-items. The cultural and evaluators characteristics sub-items were then measured using an ordinal scale with 9 points. In conclusion, the means and the medians of all the evaluators and cultural characteristics are reasonably high (Table 1).

We also note that Table 2 indicates the respondents have a strong perception of the characteristics of culture in the work place. The evaluators' intrinsic quality and cultural characteristics show strong correlation $p < 0.001$ (see Table 2). It is also interesting to see the correlation between evaluators' characteristics and cultural characteristics. Some evaluators and cultural characteristics appear to be more important than others. These characteristics were believed by the respondents to be significant attributes in most cases. This is probably, because these attributes are part

Table 1 Evaluator's characteristics

Evaluator's characteristics	Mean	Median	Minimum	Maximum	Std. Deviation
Competence	7.29	7.0	6.0	9.0	1.12
Objectivity	7.00	7.0	6.0	9.0	0.93
Knowledge	7.17	7.0	5.0	9.0	1.37
Decision making	7.17	7.0	5.0	9.0	1.17
Active thinking	7.50	7.0	7.0	9.0	0.72
Valid N = 24 (Average list wise)	7.23				

Table 2 Correlations between evaluator's characteristics and evaluator's culture

	Timeliness	Impersonal relationship	Job stability	Intercultural intelligence	Reward mechanism	Communication	Team experience
Competence	0.822**	0.549**	0.141	−0.11	0.09	0.013	0.073
	0	0.006	0.511	0.609	0.676	0.952	0.735
Objectivity	0.824**	0.616**	0.356	0.203	0.209	0.039	0.035
	0	0.001	0.088	0.342	0.326	0.855	0.872
Knowledge	0.197	0.425*	0.841**	0.329	−0.076	−0.246	−0.275
	0.356	0.039	0	0.117	0.725	0.247	0.193
Decision making	0.817**	0.634**	0.217	0.225	0.089	0.05	0.003
	0	0.001	0.309	0.291	0.678	0.817	0.988
Active thinking	0.532**	0.542**	0.460*	−0.044	−0.054	−0.204	−0.202
	0.007	0.006	0.024	0.84	0.802	0.34	0.345

*Correlation is significant at the 0.05 level (2-tailed)
**Correlation is significant at the 0.01 level (2-tailed)

of the individuals' characters, shaped by interaction with others and by life experience in the community. The low correlation of the characteristic Team Experience with the other characteristics may be because it may not be a defining characteristic of culture. Seven cultural characteristics were identified [7].

The evaluators' background (language, values, and beliefs) can affect teams' motivation. Sixth, communication skills need to be taken into account and should include all verbal and nonverbal behavior between people, including language, thoughts and feelings, problem solving, and learning by using communication channels and dialogue. As a result, a lack of communication leads to a loss of time and productivity. Finally, team experience is related to the task under investigation, the level of confidence and hard work. Personal experience affects teamwork, as each individual comes with a different set of skills gained from a different organizational culture.

To illustrate our model, a survey to assess culture was carried out to collect data on each of the four characteristics (C, S, O, Q as defined above). Each characteristic was categorized into sub-items that were rated by the respondents on a nine-point type scale from 1 to 9, where 1 means "Not Influential" and 9 means "Highly Influential". Table 3 summarize the characteristics and their sub-items. Then, for each of the main characteristics the average of the sub-items rating was calculated.

We also identified four levels of specialized skills sets. Software quality and management provides a valid rationale for considerations of knowledge and skills. Conformance is an important and initial task since it could ensure better quality of system development performance and incorporation of some keys for a particular system. It was defined from the viewpoint of a system with available capability and compatibility. System complexity is split up according to application types-supporting and core applications.

We also identified three organizational types. Project-oriented where project manager has the highest power in making decisions), Matrix in where project manager

Table 3 Evaluators' characteristic values

ID-{Q}	Optimum			Evaluator A (Requried)			Evaluator Xi		
Evaluator characterstics	Attrib	Level	Weight	Attrib	Level	Weight	Attrib	Level	Weight
Competences	9	3	0.20	7.29	3	0.20	6.5	3	0.20
Objectivity	9	3	0.20	7.00	3	0.20	5.3	2	0.20
Knowledgeably	9	3	0.20	7.17	3	0.20	6.3	3	0.20
Decision Making	9	3	0.20	7.17	3	0.20	6.5	3	0.20
Active thinking	9	3	0.20	7.50	3	0.20	6.6	3	0.20
Evaluator capability	**v(q)**	**w(q)**		**v(q)**	**w(q)**		**v(q)**	**w(q)**	
v(q1)	0.2	1		0.162	0.810		0.14	0.72	
v(q2)	0.2	1		0.156	0.778		0.05	0.26	
v(q3)	0.2	1		0.159	0.797		0.14	0.70	
v(q4)	0.2	1		0.159	0.797		0.14	0.72	
v(q5)	0.2	1		0.167	0.833		0.15	0.73	
Total v(q)	**1.00**			**0.80**			**0.63**		
Length	5.0			5.0			5.0		
p(q)-compatibility length	1.0			1.00			1.00		
w(q)-compatiblity weight	1.0			0.33			0.07		
w(q(1/4)) comp.weight avail.	1.0			0.43			0.27		

has moderate power in making decisions, Functional where project manager has the lowest level of power in making decisions. We identified five evaluators' characteristics: (1) level of competence of evaluators; (2) objectivity; (3) knowledgeability; (4) decision-making; and (5) active thinking.

These attributes associate each characteristic with a specific weight of significance, can also influence the software evaluation process The only constant inputs to this ranking system are the weights that we assign to each characteristic. These weights were obtained empirically by conducting surveys and interviews with managers and employees from the industry. The evaluators were compared against average case evaluators, derived from the values that summarize the characteristic under investigation.

5 Illustrative Example

The evaluator's characteristics are described by 5 factors for capability and compatibility values. All factors have equal weights (0.2). Then, a suitable evaluator will be selected as a prime candidate from possible evaluators, based on his capability and compatibility aspects.

We assume that an evaluator (A) was obtained from the average observed cases of evaluator's attributes. The corresponding data values were collected and considered as average cases. An ordinary project evaluator (X) was compared with the 'optimum' project evaluator. The required case helps in assessing new cases. A new case is evaluated and compared with a given required case. The candidate is assessed using the obtained values for the attributes. The measured capabilities and compatibilities are used to determine how close the candidate's case is to the required case (average case). The decision is based on closest weight. For example, the evaluator's profile capability for required case (A) is:

$$V_{(q1)} = 0.20\left(\frac{7.29}{9}\right)\left(\frac{3}{3}\right)^2 = 0.162$$

The levels of scale are: 1 for low, 2 for nominal and 3 for high. The total capability $V_{(q)}$ is the sum of all $V_{(qi)}$, where i is the number for factors. The length of evaluator's factors is the number available required by the evaluator's profile. Table 3 shows a summary of evaluator's capability—optimum, required and possible evaluator $x_{(i)}$. The capabilities of the required evaluator (A) and possible evaluator $x_{(i)}$ are computed as 0.80 and 0.63 respectively. The possible evaluator $x_{(i)}$ shows lower capability profile based on selected values. The compatibility is lower for possible evaluator $x_{(i)}$ than the required average case.

The compatibilities weight was computed using all evaluator's characteristics. The profile for required evaluator (A) shows 0.33 and capability weight available is 0.43. The model described the foundation for devising a simple algorithm to select the most suitable software evaluator's profile that matches the project manager's optimal requirements.

The compatibility length (p_q) is the ratio of available factors over the required factors. For example, the p_q for possible evaluator $x_{(i)}$ is: 5/5

$$p_q = \frac{\# available\ factors}{\# required\ factors} \leqslant 1$$

The compatibility weight (q) is an integrated quality in the evaluation system. For example, the ω_q for an average evaluator (A) is:

$$W_q = \prod_{j=1}^{m_j} \left(\frac{\varepsilon_j}{\varepsilon_j^{(())}} \right) \left(\frac{v_j}{v_j^{(())}} \right)^2 = 0.810 * 0.778 * 0.797 * 0.797 * 0.833 = 0.33$$

are the required factors capabilities. $W_{(l(1 \div 4))} = 0.43$ is the available factors capabilities. Therefore, only four factors were satisfied.

Results, the evaluation process is confined to average case evaluator and based on their values, we conclude that the evaluator (X) should not be selected because the value of his profile is the lower than required (0.63 < 0.80).

Finally, the value of the overall profile (e_{v}) is defined by the mean values of its comprising characteristics:

$$e_{\mathrm{v}} = \sum_{i=1}^{i=n} v_{i/n}$$

An illustrative scenario of the organization outlines the capability and compatibility for each evaluator's characteristic, for both the experience and competence criteria. In addition, each characteristic is associated with a special weight of significance. In light of this evaluator's profile, we need to select the best matching evaluator among candidates. Using the selection algorithm, we can reduce the candidates set by factoring out the evaluators who does not meet at least one required characteristic in this profile. We excluded the possible candidate $x_{(i)}$ because of his low level of competence, does not meet the objectivity criteria and level required for evaluator's capability. Therefore, the competition is based upon the evaluator's profile values.

To conclude, the possible evaluator $x_{(i)}$ is not selected because the value required in the survey findings for each average case (A) is much higher. This approach allows the selection of the most suitable candidates for proposed possible evaluators, conducted in a formal way.

6 CASE Tools, Techniques, and Environments

CASE tools have been adopted with the perspective of enhancing software quality and productivity. Nowadays, in software development industry, a large number of CASE tools have been produced to perform tasks to adapt them with automated support and standardization. However, most of software development teams use CASE tools that are assembled to adopt new CASE tools without establishing formal evaluation criteria over time [15, 17]. These components proposed evaluators profile as a case tools in software evaluations.

The biggest benefit of using front-end CASE tools is in software quality, not necessarily productivity. The ability to have the computer check your design for errors and/or inconsistencies is a real step forward, since preventing bugs is highly preferable to searching for them during testing. The real productivity progression will come from increased reuse of existing code and design components. The activity-based classifications of software CASE tools, shows, not only visual aids to process phases but supports the software production process. It is important for evaluators to base software quality on some CASE tools. Many of these tools are relevant to all/some software development phases and it is also critical to support testing and software quality.

We suggest that an appropriate software evaluation is acquired when an evaluator's cultural attributes are incorporated in the evaluation process. Our work focuses on considering a formal mean when selecting the most appropriate evaluator in light of culture aspects. The work showed using an example whose aspects are empirically driven.

7 Conclusion

Evaluation constitutes a crucial activity at our disposal to appraise the validity of the tools used to test software appropriateness in order to improve the product at every phase in the software life cycle. Given the globalization of software development and deployment, the profile of software engineers cannot be limited to just technical skills. Profile Theory was used to develop a model that captures the essential technical and cultural characteristics of contextually-effective software evaluators. These evaluator-defining characteristics include cultural, organizational, technical, and individual attributes and relationships among them. The incorporation of our model in software development as a CASE tool addresses directly the issues of evaluation in a global environment.

References

1. Krishna S, Sahay S, Walsham G (2004) Managing cross-cultural issues in global software outsourcing. Commun ACM 47(4):62–66
2. Cater-Steel A, Toleman M (2008) The impact of national culture on software engineering practices. Int J Technol Policy Manag Indersci Publishers 8(1):76–80
3. Liang T, Liu C, Lin T, Lin B (2007) Effect of team diversity on software project performance. Ind Manag Data Syst 107(5). Emerald Group Publishing Limited, pp 636–653. www.emeraldinsight.com/0263-5577.htm
4. Clemmensen T, Goyal S (2005) Cross cultural usability testing. Working paper, Copenhagen Business School, Department of Informatics, HCI research group, p 20
5. Hofstede G, Hofstede J (2005) Cultures and organizations: software of the mind revised and expanded, 2nd edn. McGraw-Hill, New York
6. Hamdan K (2008) An investigation into software estimation methods, PhD Thesis Sunderland, UK
7. Hamdan K, Smith P, Plekhanova V (2012) Leadership and Cultural Issues: Evaluation and Measurement in the Context of Software Development. Int J Inf Educ Technol. https://doi.org/10.7763/IJIET.V2.85
8. Schein E (2004) Organizational culture and leadership, 3rd edn. Joddry-Bass, John Wiley & Sons, pp 129–225
9. Hertzum M, Jacobsen N (2001) The evaluator effect: a chilling fact about usability evaluation methods. Int J Hum-Comput Interact 13(4):421–443
10. Gillies A, Smith P (1994) CASE usage in the UK, 1991. In: Managing software engineering. Springer US, pp 42–50
11. Crosby P (1980) Quality is free. Mentor Books, New York
12. Fitzpatrick R, Smith P, O'Shea B (2004) Software quality revisited, proceedings of the software measurement. European Forum (SMEF 2004, Rome), Istituto di Ricerca Internazionale S.r.l., Milan, Italy, p 307/315, ISBN 88-86674-33-3
13. Roper R, Smith P (1988) A specification-based functional testing method for JSP designed programs. Inf Softw Technol 30(2):89–98
14. Roper M, Smith P (1987) A structural testing method for JSP designed programs. Softw: Pract Exp 17(2):135–157
15. Boehm B, Baik J (2000) Empirical analysis of case tool effects on software development effort
16. Jackson M, Crouch S, Baxter R (2016) Software evaluation: criteria-based assessment, software evaluation: criteria-based. http://www.software.ac.uk/software-evaluation-guide
17. Majchrzak T (2012) Improving software testing technical and organizational developments. Briefs in information systems, 1st edn. Springer
18. Dubey S, Sharma D (2015) Software quality appraisal using multi-criteria decision approach. Int J Inf Eng Electron Bus 7(2):8–13
19. Futrell R, Shafer D, Shafer L (2002) Quality software project management. Software Quality Institute
20. Plekhanova V (2000) Applications of the profile theory to software engineering and knowledge engineering, Chicago, pp 133–141
21. Plekhanova V (2000) On the compatibility of contemporary project management tools with software project management, information systems management, Orlando

Transforming JavaScript-Based Web Application to Cross-Platform Desktop with Electron

Kitti Kredpattanakul and Yachai Limpiyakorn

Abstract Over the years, various ways emerge for evolving the existing applications towards a shared codebase. Among several, Electron is a well-known framework for web developers to build cross-platform desktop applications using familiar web technologies, such as HTML, CSS, and JavaScript. This paper thus presents an approach for transforming web applications created with JavaScript to desktop applications that can run on Windows, MacOS, and Linux. The output desktop application would remain the old set of source code for further development.

Keywords Cross-platform desktop · Web application · Electron framework JavaScript

1 Introduction

Today, the notable JavaScript frameworks or libraries, such as ReactJS and AngularJS, are popular for web application development. The advantage of web applications is its capability to work across multiple types of platforms or operating environments via web browsers, whereas a desktop application is hard to manage on various sets of source code and resources for the operability on more than one platform with identical (or nearly identical) functionality. However, the web application needs abilities of desktop such as native menus and notification.

The advent of the Electron framework lets developers write cross-platform desktop applications using JavaScript, HTML and CSS. Electron [1] is an open-source framework created and maintained by GitHub. It allows for the development of desktop GUI applications using front and back end components originally developed for web applications: If you can build a website, you can build a cross-platform desktop

K. Kredpattanakul · Y. Limpiyakorn (✉)
Department of Computer Engineering, Chulalongkorn University, 10330 Bangkok, Thailand
e-mail: Yachai.L@chula.ac.th

K. Kredpattanakul
e-mail: Kitti.Kre@student.chula.ac.th

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_56

application with Electron that can run on Windows, MacOS and Linux. Development with Electron framework is easy since web application developers do not have to learn new languages for specific operating systems. Therefore, it would reduce the development time as well as the cost of production. This paper thus presents an approach for transforming web applications created by ReactJS to cross-platform desktop applications built with the installation file. The output desktop application would remain the old set of source code for further development.

2 Background

2.1 JavaScript-Based Web Application

JavaScript has become the popular language for implementing modern web applications. The language is commonly used for loading core functionality and user interfaces to the client side, also implementing real-time applications on the server-side through environments such as Node.js [2]. The traditional server-side templating is moving to JavaScript templating with many new JavaScript frameworks that drives user interfaces become more mobile as they are given more hardware access through JavaScript APIs. JavaScript is now getting objects and data from RESTful or WebSocket connections, which are automatically bound to the user interface through a client-side, JavaScript framework [3].

2.2 Electron Framework

Electron is an open source framework for creating native applications with web technologies such as JavaScript, HTML, and CSS [1]. Electron works by combining the chromium content framework and Node.js together in a single framework [4]. Electron can produce a desktop application which can run on Windows, MacOS and Linux. The old style development of desktop applications uses various development teams for different operating systems, resulting in multiple versions of source code, each of which is associated with an individual operating system. Whereas the desktop applications built with Electron would ensure more maintainability by using one set of source code, also reduce the cost and time of development.

3 Research Methodology

The proposed approach of transforming a JavaScript-based web application to a cross-platform desktop application is depicted in Fig. 1. One of the web application

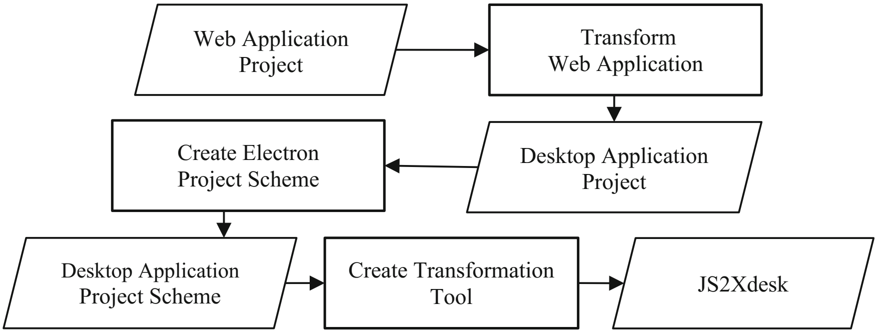


Fig. 1 Overview of transformation tool construction

development projects from official ReactJS web site [5], Calculator, is selected to demonstrate the procedure.

Based on the presented method, the input of web application being transformed is required to comply with the following requirements:

- The web application must have the command to build static files for production.
- The web application must work well before transforming.
- The web application must have the package.json file.
- The user interface must be implemented with ReactJS.

The output is the transformation tool called JS2Xdesk, which is the executable invoked via the command line.

3.1 Transform Web Application

Initially, the example run as web application, Calculator (Fig. 2), is selected from the ReactJS official web site [5], which also provides the source code for developers as open source project. The source is reliable to transform the web application project into desktop. The associated structure of ReactJS application project is shown in Fig. 3. Referring to the Electron official web site [6], the selected structure of Electron desktop project is illustrated in Fig. 4.

The new development project created by merging the Calculator React project files (Fig. 3) with Electron desktop project files (Fig. 4) is shown in Fig. 5. The dependencies property of a module's package.json is where dependencies—the *other* modules that *this* module uses—are defined. [7]. The scripts and dependencies from the Calculator projects are also merged as shown in Fig. 6. The static.js and.css files from the Calculator React project is put to .html file of new project as shown in Fig. 7.

Once the manual preparation of package.json and index.html files has been completed, the execution of the start command of the new project built will generate the

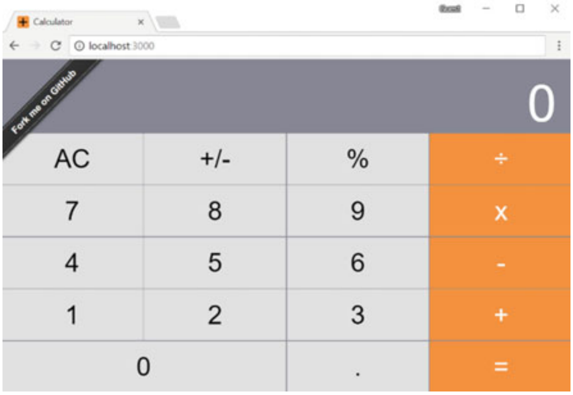
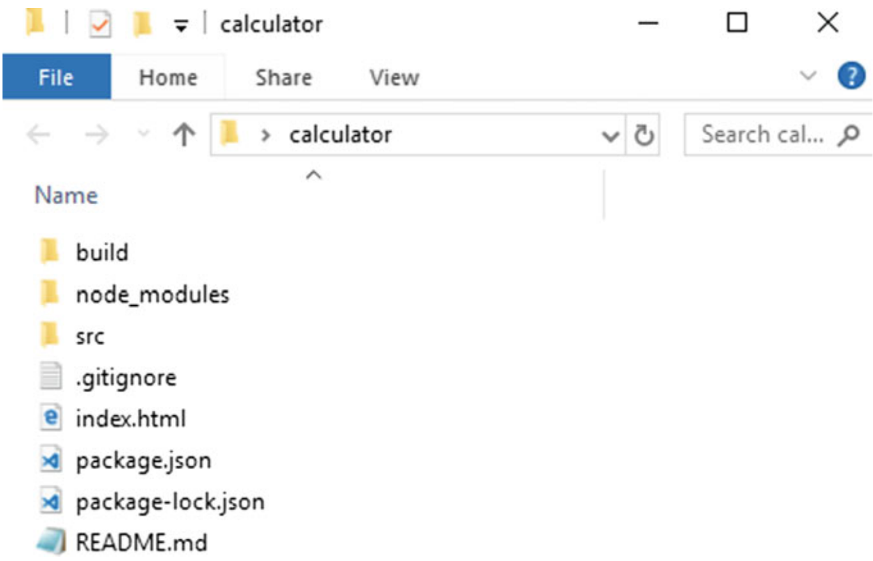


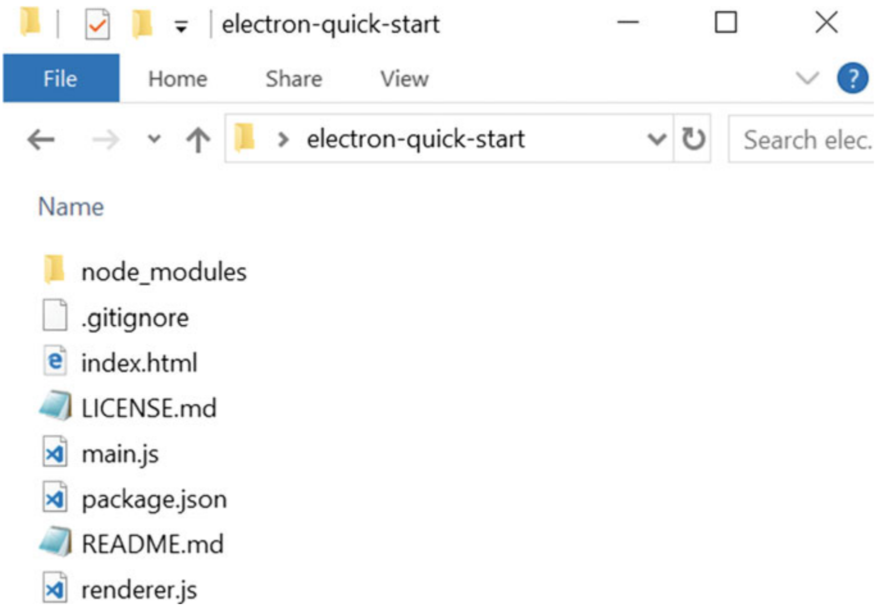
Fig. 2 Example Calculator web application



build
node_modules
src
.gitignore
index.html
package.json
package-lock.json
README.md

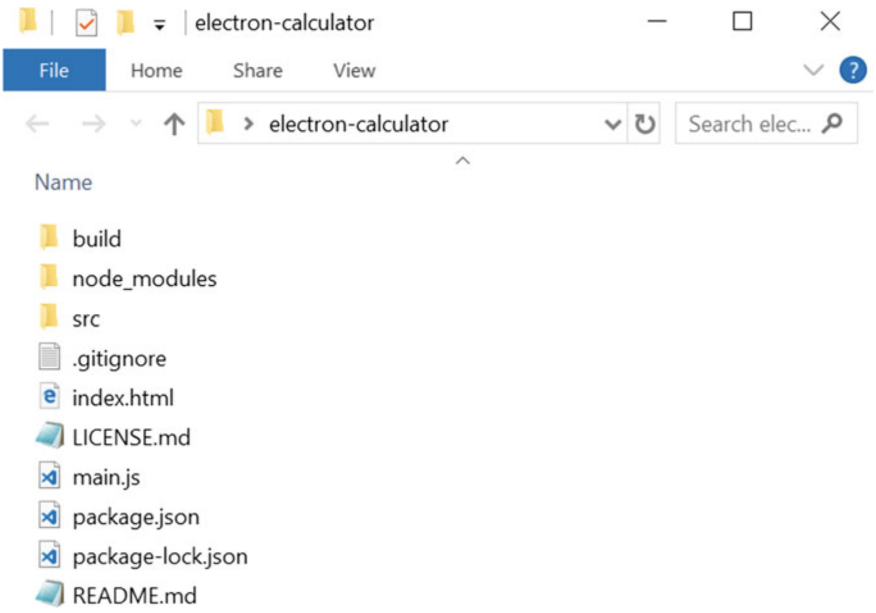
Fig. 3 Structure of Calculator web application project

output as shown in Fig. 8, which is rendered on Windows. The style of the menu bar and title bar would be varied on different platforms due to the platform's interfaces. However, the screen layouts and business logics are still retained on different platforms, such as MacOS and Linux.



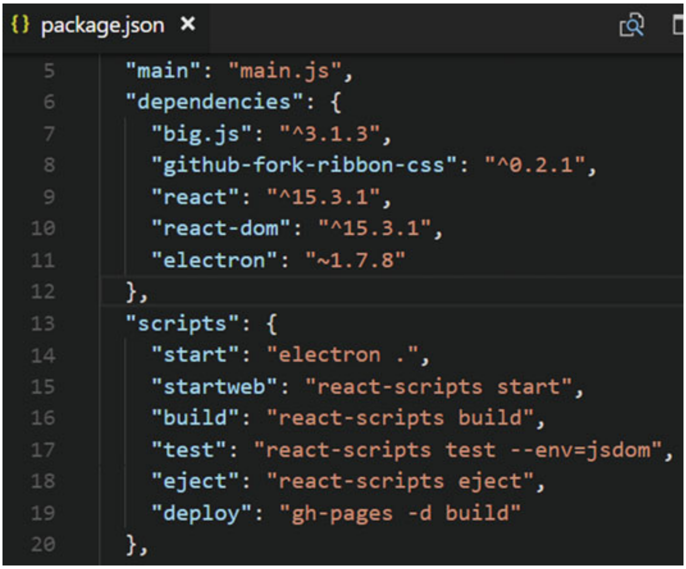
node_modules
.gitignore
index.html
LICENSE.md
main.js
package.json
README.md
renderer.js

Fig. 4 Structure of Electron desktop application project



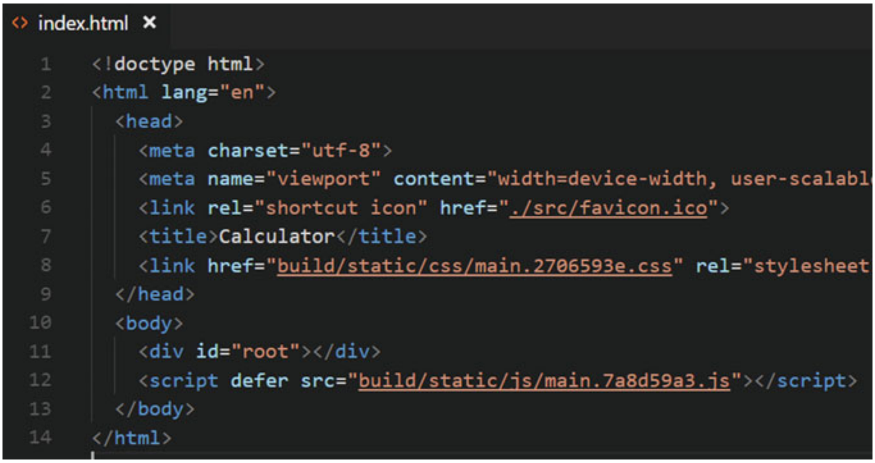
build
node_modules
src
.gitignore
index.html
LICENSE.md
main.js
package.json
package-lock.json
README.md

Fig. 5 Merged structure of web project and desktop project



```
"main": "main.js",
"dependencies": {
  "big.js": "^3.1.3",
  "github-fork-ribbon-css": "^0.2.1",
  "react": "^15.3.1",
  "react-dom": "^15.3.1",
  "electron": "~1.7.8"
},
"scripts": {
  "start": "electron .",
  "startweb": "react-scripts start",
  "build": "react-scripts build",
  "test": "react-scripts test --env=jsdom",
  "eject": "react-scripts eject",
  "deploy": "gh-pages -d build"
},
```

Fig. 6 Scripts and dependencies after merge



```
<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, user-scalabl
    <link rel="shortcut icon" href="./src/favicon.ico">
    <title>Calculator</title>
    <link href="build/static/css/main.2706593e.css" rel="stylesheet
  </head>
  <body>
    <div id="root"></div>
    <script defer src="build/static/js/main.7a8d59a3.js"></script>
  </body>
</html>
```

Fig. 7 HTML file of Electron desktop application project

3.2 Create Electron Project Scheme

Figure 9 illustrates the build directory containing static app.css and app.js from the web application project. The node_modules directory contains dependency modules from both web application and Electron project. The src directory contains web application source code. The index.html file is the main .html file that contains the

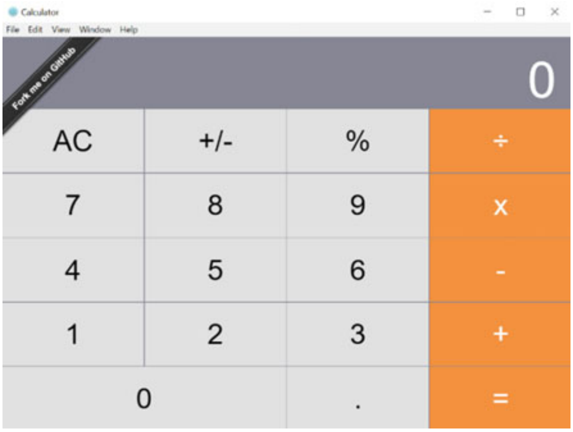
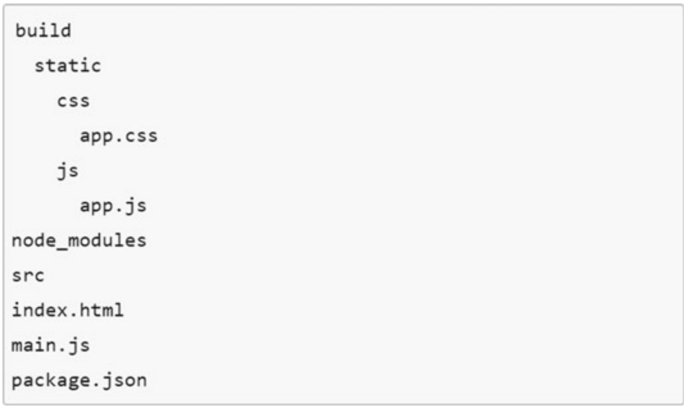


Fig. 8 Transformation output of electron desktop application on Windows



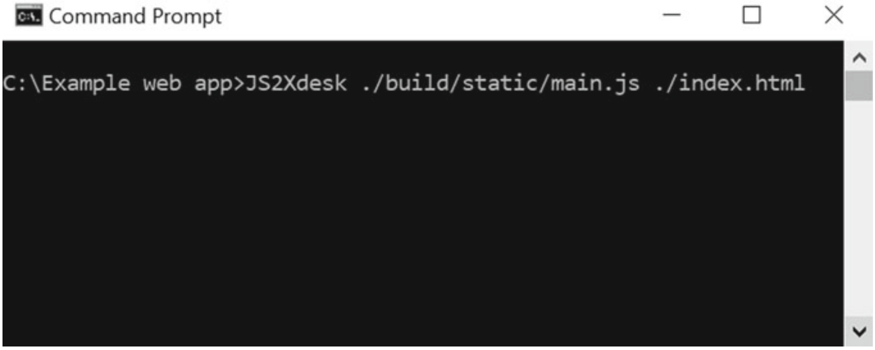
```
build
  static
    css
      app.css
    js
      app.js
node_modules
src
index.html
main.js
package.json
```

Fig. 9 Electron project scheme created in folder build

path of app.css and app.js. The main.js from Electron project contains the path of index.html file. And the package.json is the merged script from web application project and Electron project.

3.3 *Create Transformation Tool*

The significant point to understand in Electron framework is its own architecture [8]. There are two types of processes available in Electron: 1. main (main.js) process, and 2. renderer (app.js) process. The main.js contains the script to run in the main process that includes the resolution of the application and the path of .html file, in addition to



```
C:\Example web app>JS2Xdesk ./build/static/main.js ./index.html
```

Fig. 10 Display of arguments and command invoked for transformation

the application menu bar that the transformation can provide only the default function such as toggle a full screen, minimize, and close the window. The main process creates a web pages by creating an application window. Each application window runs the web page in its own renderer process which is responsible for loading web pages to display the GUI. In this work, the app.js is the application container that dynamically reaches every page in the application and already included in the .html file with the compiled.js file.

The presented method requires the construction of the folder named *build* containing the Electron project scheme as shown in Fig. 9. The Node.js technology and libraries from npm [9] are then used for creating the transformation tool, so called JS2Xdesk. Figure 10 illustrates the screen shot displaying the command for executing the transformation program. The command format is defined as follows:

$$\text{JS2Xdesk}[main.js\ path][html\ file\ path]$$

Note that the new version of the transformation tool, JS2Xdesk, needs to be re-created whenever the version of Node.JS in Electron is not the last updated as appearing in the web application.

Another web application from the official ReactJS web site [5] is selected for testing the correctness of the created tool. Figure 11 illustrates the valid result of transformation to desktop application on Windows.

4 Conclusion and Future Work

The desktop application usages will continue to increase for business and life due to ease of production with helpful technology like Electron framework. This paper, therefore, presents an approach to transform web applications to cross-platform desktop apps with Electron framework. The completed application produced in web

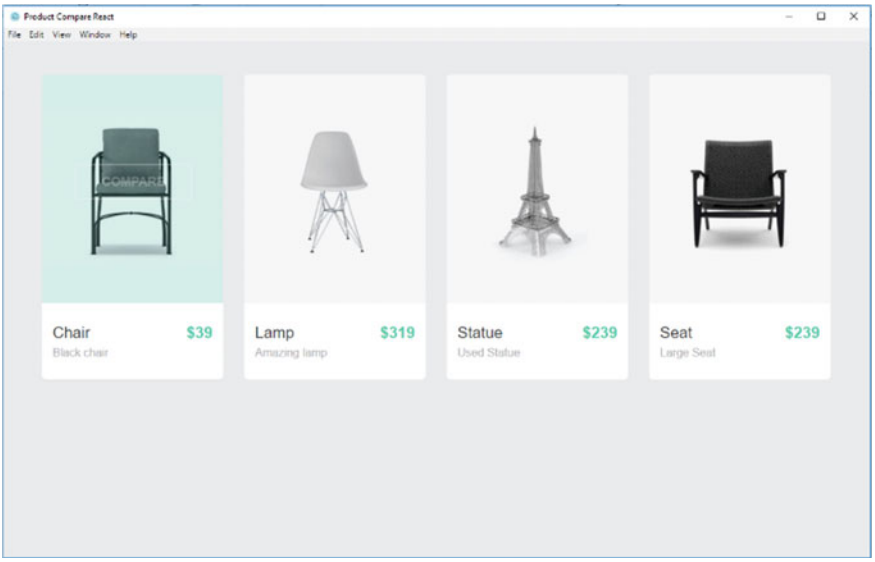


Fig. 11 Valid test result of transformation with JS2Xdesk

application technology is required when being converted to the cross-platform desktop application without any change of business logic source code. The proposed approach is beneficial as an alternative for reducing the resources consumed when transforming the application projects developed with web technologies to the desktop applications that can operate on multiple platforms.

References

1. Electron. https://electronjs.org. Accessed 21 Dec 2017
2. Node.js. https://nodejs.org/en. Accessed 21 Dec 2017
3. Hales W (2012) HTML5 and javaScript web apps. O'Reilly Media, California
4. Jasim M (2017) Building cross-platform desktop applications with electron. Packt Publishing, Birmingham
5. ReactJS example project. https://reactjs.org/community/examples.html. Accessed 10 Dec 2017
6. Electron example project. https://github.com/electron/electron-quick-start. Accessed 10 Dec 2017
7. The basics of package.json in Node.js and npm. https://nodesource.com/blog/the-basics-of-package-json-in-node-js-and-npm. Accessed 12 Apr 2018
8. Electron application architecture. https://electronjs.org/docs/tutorial/application-architecture. Accessed 4 Apr 2018
9. npm. https://www.npmjs.com. Accessed 20 Dec 2017

Degree of Similarity of Root Trees

Jiri Sebek, Petr Vondrus and Tomas Cerny

Abstract Adaptive User Interfaces (UI) provide better user experience as users a receive personalized presentation. These UIs heavily rely on contextual data. Context helps the application to recognize user needs and thus adjust the UI. First time user receives a generalized experience; however, as the user uses the application more often it gathers lots of contextual data, such as the history of actions. This allows to statistically classify user in a user cluster and based on that adapt the UI presentation. This paper considers methods to find a measure of similarity of graphs to support adaptive UIs. To achieve this, it considers rooted trees. It states known approaches, which could be used for calculation of this measure. It focuses on the Simhash algorithm and describes its implementation in the SimCom experimental comparative application. Its results show that Simhash can be used for comparing the rooted trees. The main aim of this paper is to show novel view on how to use graph algorithms and clustering of trees into adaptive application structure.

Keywords Graph algorithms · Digraph · Rooted tree · Similarity · Simhash
Context-aware user interface

1 Introduction

To attract users, applications provide personalization and adaptability features in User Interface (UI). The challenge is how to provide an efficient and most suitable variation of UI assembled specifically to a particular user and his/her needs. Applications can

J. Sebek (✉) · P. Vondrus
Department of Computer Science, Czech Technical University, Prague, Czech Republic
e-mail: sebekji1@fel.cvut.cz

P. Vondrus
e-mail: vondrp11@fel.cvut.cz

T. Cerny
Department of Computer Science, Baylor University, Waco, TX, USA
e-mail: tomas.cerny@baylor.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_57

adapt differently based on type of context they get [12]. For instance, [13, 15] suggests a framework to adapt UI menu based on history of user actions. Every application has some kind of UI element that divide application. This structure element (menu) we can represent as root tree. Reference [12] work adapt UI elements based on sensors data. Another interesting information is user emotion which can be important information for adaptation [14]. All of these adaptations have same basic idea. We can improve these approaches because we know some users will have similar UI (e.g. teachers, students of same branch). If they create same user cluster we can no longer wait for all context data and adapt UI based on cluster information. We can accelerate this process. This paper addresses one challenge in adaptable UIs related to graph similarities. Specifically, it considers possibilities of algorithmic comparison of two graphs of the root category tree to design and implement different methods for calculating the measure their similarities. This resolves the problem of first gathering a lot of information about the user before we can adapt its UI. Graph theory as a component of modern discrete mathematics finds broad application in the present computer science. The introduction to this issue is explained in detail in a wide range of publications, such as [2, 6, 8, 16]. This paper is organized as follows. Section 2 describes the background for similarity and its calculation. Sections 3 introduces method Simhash. Section 4 contains Implementation with experiments and shows results. Section 5 presents conclusions.

2 Similarity and Its Calculation

In algorithmic comparison of two data objects, there is no problem finding a match or the difference, but there is problem to relatively quickly determine the degree of this difference. Conventional methods comparisons are very inefficient and slow in the case of large objects. In addition, in in many cases, with sufficient precision to estimate the degree of similarity. Most modern algorithms [5, 20] perform this general procedure when calculating the degree of similarity: 1. Transform an object into a suitable metric space [7] 2. Calculate their distance in this space using the selected metric The goal is to express the similarity of two data objects A, B with such a function

$$\mathrm{sim}(\mathrm{A}, \mathrm{B}) \in \langle 0 - 1 \rangle \tag{1}$$

so that its functional value increases with the data object match. Here are some frequently used metrics to determine the degree of similarity: We can divide these metrics into two main groups: In the R n space and in discrete metric space. Into first group belongs metrics like Euclidean distance or Manhattan distance. The second group contains Hamming's metric and Levenshtein distance. We could also use the existing algorithm methods to compare rooted trees based on editing distance (Selkow's algorithm [17], Chawathe's algorithm [4], Tekli's algorithm [19]) Another method, that calculate similarity, is the simhash method. It is elegant way to calculate the degree of similarity of arbitrarily large data objects. It is based on Simhash function. The Simhash function belongs to the category of Locality-sensitive hashing (LSH) hash functions that they reduce the dimensionality of input data and unlike cryptographic hash functions tries to maximize the number of collisions for similar templates. The next subsections are dedicated for this method.

2.1 Method Simhash

The author of simhash method is Moses Charikar [3] and it was published in 2002. It uses hash function and Hamming distance. The hash function [10] is such a non invertible a representation of the key to a relatively large set of U assigns a fingerprint from a relatively small set H. In practice, hash functions have a wide range of applications. They are used for quick search using hash tables, data integrity detection or cryptography.

Hamming distance [1] is metric in the text strings space and represents the number of places in which they are two objects of the same data structure different or the number of corrections that are needed to change one object into another one. For two objects x and y with length k it can be expressed by a relationship:

$$\mathrm{DH} = \mathrm{n} \sum \mathrm{i} = 1|\mathrm{xi} - \mathrm{yi}| \tag{2}$$

Common, especially cryptographic, hash functions often break the correlation input data and fingerprint using the avalanche effect, when a small change of input (several bits) causes a large change of the fingerprint (one-half bits). For a set of very similar templates, then creates a set of completely different fingerprints. Instead, Simhash belongs to the group perceptual hash [21] which produces very similar fingerprints for very similar original sources. For non-trivial original sources, it reduces asymptotic complexity because of its usage it is not necessary to compare two original sources by the iterative walkthrough of their internal structures. It is faster (asymptotically) count their Simhash fingerprints and compare.

2.2 The Simhash Walkthrough

This section contains the the steps of the algorithm. It is explained on simple example. Let's have two X and Y short strings

$$\begin{aligned} X &= \text{``}Wehavefirstoriginaltextthatweneedtocompare.'' \\ Y &= \text{``}Wehavesecondoriginaltextthatwewilltocompare.'' \end{aligned} \tag{3}$$

Table 1 Table of MD5 hash from the Tx set

Token	MD5 [hex]	MD5 [bin]
We have first original	5599477AF16D0C8F15736FD045C6D8C1	10101011…11000001
	1869E7D1AC1F6274DB315A92F22A9EFC	11000011…11111100
	EB260E9AE827821BECEEED4104F0AD89	11101011…10001001
	88FA9F694690E11239096536CCF2702B	10001000…00101011
…	…	…

Table 2 Table of MD5 hash from the T_y set

Token	MD5 [hex]	MD5 [bin]
We have second original	5599477AF16D0C8F15736FD045C6D8C1	10101011…11000001
	1869E7D1AC1F6274DB315A92F22A9EFC	11000011…11111100
	59D0D19FC45CA69230D858F60A5557F8	10110011…11111000
	88FA9F694690E11239096536CCF2702B	10001000…00101011
…	…	…

We see that the sentences are similar. The difference is only in 2 of 7 words, respectively. at 12 from 44 characters. Subjectively, we expect the resulting similarity rate to be high. So we assume sim (X, Y) ∈⟨ 0. 70 − 0. 80 ⟩

Determining the length of the fingerprint. The length of the fingerprint (algorithm output) is equal to the length of the hash function used (see step 3). Usually it is 32, 64 or 128 bits. This option increases accuracy algorithm, but it reduces its speed. Using cryptographic hash function MD5 [11] The fact that MD5 has been compromised for the purpose Of course, this method does not matter., the 128-bit print length will be.

$$\mathrm{n} = 128 \tag{4}$$

Splitting the original The original object is divided into elementary data entities (also called tokens), which become elements of a set of tokens T. For original X and Y the sets of words $\mathrm{T_X}$ and $\mathrm{T_Y}$ are created. Punctuation marks are here ignored.

$$\begin{aligned} &\mathrm{T_X} ='' \text{ We}''\text{,}''\text{ have}''\text{,}''\text{ first }''\text{,}''\text{ original}''\text{,}''\text{ text}''\text{,}''\text{ that}''\text{,}''\text{ we}''\text{,}''\text{ need}''\text{, } \ldots \\ &\mathrm{T_Y} ='' \text{ We}''\text{,}''\text{ have}''\text{,}''\text{ second}''\text{,}''\text{ original}''\text{,}''\text{ text}''\text{,}''\text{ that}''\text{,}''\text{ we}''\text{,}''\text{ will}''\text{, } \ldots \end{aligned} \tag{5}$$

Using hash function Each element t of the set of tokens T has the chosen hash function. This creates a set of fingerprints H. For token sets $\mathrm{T_X}$ and $\mathrm{T_Y}$ there will be created the set of fingerprints $\mathrm{H_X}$ and $\mathrm{H_Y}$ (Tables 1 and 2).

$$\begin{aligned} H_X &= 559947.., 1869E7.., EB260E.., \ldots \\ H_Y &= 559947.., 1869E7.., 59D0D1.., \ldots \end{aligned} \tag{6}$$

Creating a distribution vector The array of integers A is declared with a length of n, which is initialized with zeros. The field contents are then filled in as follows:

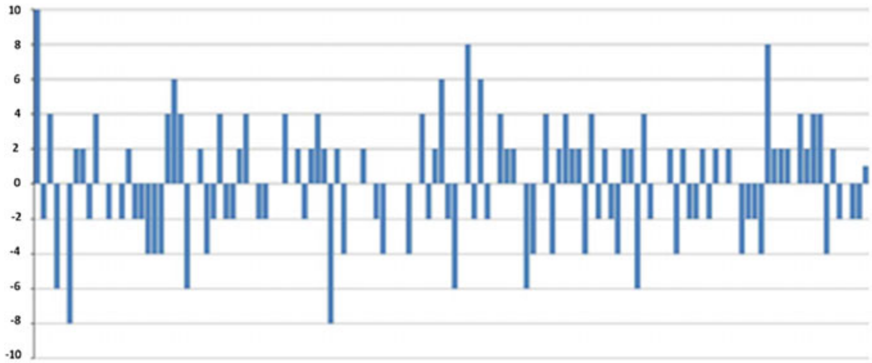


Fig. 1 Graph of the distribution vector of the origin

$$\mathrm{h} \in \mathrm{H}, \mathrm{t} \in \mathrm{T}, \mathrm{i} \in \mathrm{N}, 1 \leq \mathrm{i} \leq \mathrm{n}$$

$$A[i] = \begin{cases} A[i] - weight(t) for\, bit(h, i) = 0, \\ A[i] + weight(t) for\, bit(h, i) = 1. \end{cases} \tag{7}$$

Each token can have an integer weight corresponding to some of its properties (e.g., the frequency of occurrence), which provides the weight function weight (t). This makes it possible to increase the accuracy of the algorithm for certain types of templates (here weight (t) = 1). For the sets of fingerprints $\mathrm{H_X}$ and $\mathrm{H_Y}$, two integer fields will be created (Fig. 1)

$$\begin{aligned} A_X &= [10, -2, 4, -6, 0, -8, 2, 2, -2, 4, 0, -2, 0, -2, 2, -2\ldots] \\ A_Y &= [10, -4, 4, -4, -2, -8, 4, 4, 0, 4, 0, 0, 0, -6, 2, 0\ldots] \end{aligned} \tag{8}$$

Creation of the fingerprint It is declared bit field Q with length n, which is filled in as follows: (Fig. 2)

$$\mathrm{i} \in \mathrm{N}, 1 \leq \mathrm{i} \leq \mathrm{n}$$

$$Q[i] = \begin{cases} 1\ for\ A[i] > 0, \\ 0\ otherwise. \end{cases} \tag{9}$$

Q is the final fingerprint, respectively output of the hash function Simhash. For the $\mathrm{A_X}$ and $\mathrm{A_Y}$ distribution vectors, two bit arrays will be created $\mathrm{Q_X}$ and $\mathrm{Q_Y}$.

$$\begin{aligned} Q_X &= [1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0\ldots] \\ Q_Y &= [1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0\ldots] \\ Q_X &= 0xA3420E4982BA201629C2F53452A0F7A1 \\ Q_Y &= 0xA342060986BA712669A2B124D080F7A0 \end{aligned} \tag{10}$$

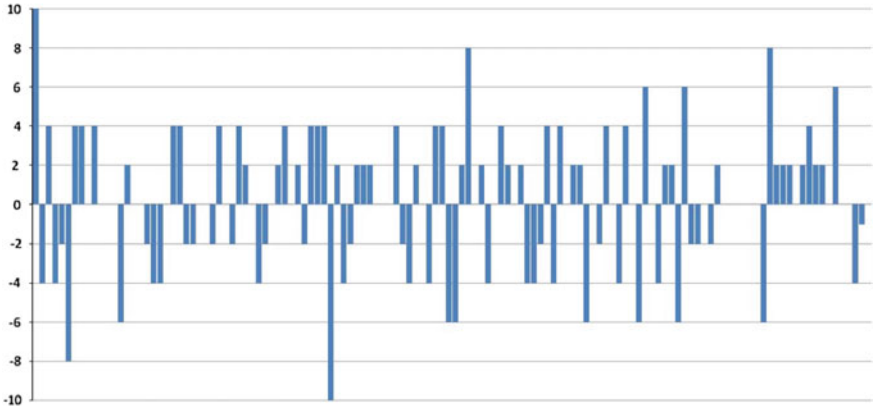


Fig. 2 Graph of the distribution vector of the origin

Calculation of the similarity degree The final step of the algorithm is to calculate the degree of similarity of the templates based on their Simhash of imprints. This is done using the Jaccard index [9].

$$sim(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{11}$$

(for X = Ø, Y = Ø define sim (X, Y) = 1), where there are used Simhash fingerprints Q X and Q Y instead of the originals X, Y

$$sim(X, Y) = \frac{|\mathrm{QX} \cap \mathrm{QY}|}{|\mathrm{QX} \cup \mathrm{QY}|} \tag{12}$$

Applying Eq. (3) then applies:

$$\begin{aligned} |Q_X \cap Q_Y| &= D_H \\ |Q_X \cup Q_Y| &= n \end{aligned} \tag{13}$$

and the resulting similarity of templates X, Y is

$$sim(X, Y) = \frac{\mathrm{DH}}{\mathrm{n}} \tag{14}$$

If we set up variables in (8), for example, using the XOR logic function we can calculate:

$$\mathrm{sim(X, Y)} = \frac{\mathrm{n} - \sum_{i-1}^{n} (Q_{Xi} \mathrm{XOR}\ Q_{Yi})}{\mathrm{n}} = 0, 86$$

The value of the resulting degree of similarity lies within the expected in- terval and captures well the estimated difference between the originals. This number tells us metric how much are two structures similar. Now if we use some representation of graphs instead of simple strings, we are able to compare data structures that represent menus or another more complex UI element. With this information, we can create user clusters of each menu. When a new user starts to use application, we know from context that the user has a role for example student, teacher, administrator, moderator so we can create better, user-friendly UI specified for his/her needs. Here we are reducing the time we need to gather historical data, or reducing the complexity of size of the context information we need.

3 Implementation

To evaluate our proposal in experiments we implement a SimCom Java FX application that uses DOT program. The implementation of the method for calculating the degree of similarity of root trees is described in Chap. 3.

Decomposition In the case of a tree, the layout options for individual tokens are already available from the graph definition itself. We can consider divisions based on vertices, edges or their appropriate combinations. We chose vertices of tree for individual tokens because unlike the edge even a single vertex is a graph.

Hash function Choosing a hash function is important in this method because it has a significant effect output Simhash fingerprints. We tried to design the layout of the respective classes open so that it is possible to gradually add different hash functions and observe their influence.

We used the Google Guava open source library[1] that offers hashing class with various hash functions. We have used murmur3.32(), sipHash24 and farmHashFingerprint64. The custom token is composed of four attributes of the class“CustomGraphVertex”.

They cover all information we need to specify location of token. Indegree attribute stores the entry level of the vertex, outdegree attribute stores the output level of the vertex. Level attribute contains the vertex depth and label is unique vertex identifier.

Traversing a tree Tree decomposition on tokens and subsequent call of hash functions is performed during a single tree pass through its Depth-first search algorithm [18].

The description of graphs The mathematical representation of a graph can be created by, for example, enumerating a set of edges E (G) or using the adjacency matrix. Instead, we used format DOT.[2] It is a plain-text language and you can see example at listing 1.1 and our application uses JGraphT library[3] for graph modeling.

[1] https://github.com/google/guava.

[2] https://graphviz.gitlab.io/pages/doc/info/lang.html.

[3] https://www.graphviz.org.

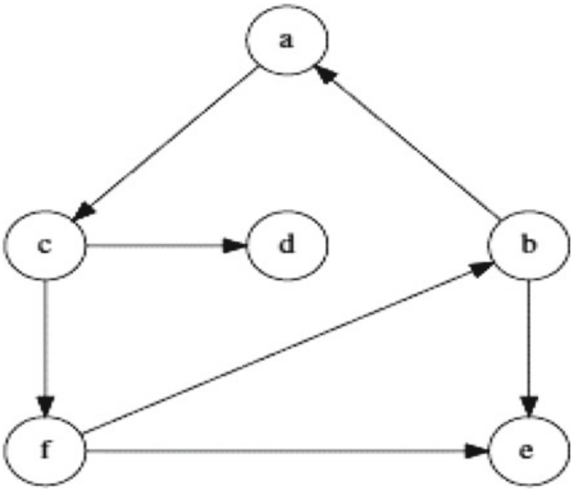


Fig. 3 This result of oriented graph description

Listing 1.1 The example of oriented graph description (Fig. 3)

```
digraph G {
    rankdir = TB;
    nodesep = 1.0;
    ranksep = "1.0 equally";
    node [shape = circle];
    a [group = g2]
     {rank = same; b[group = g1];
     c[group = g3]; d[group = g4];}
     {rank = same; e[group = g1];
     f[group = g3];}
        a -> c;c -> d;c -> f;f -> b;
        f -> e;b -> a;b -> e;
   }
```

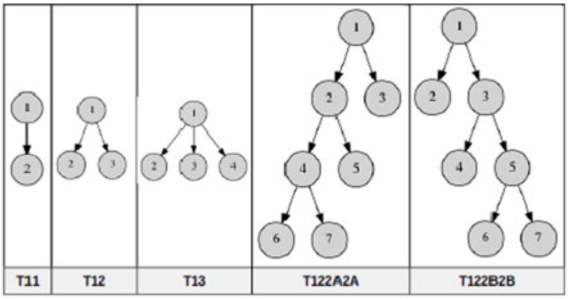
Catalog and persistence Imported charts are stored in the *catalog*, which is a dynamic data array representing the structure of graphs that the application can use.

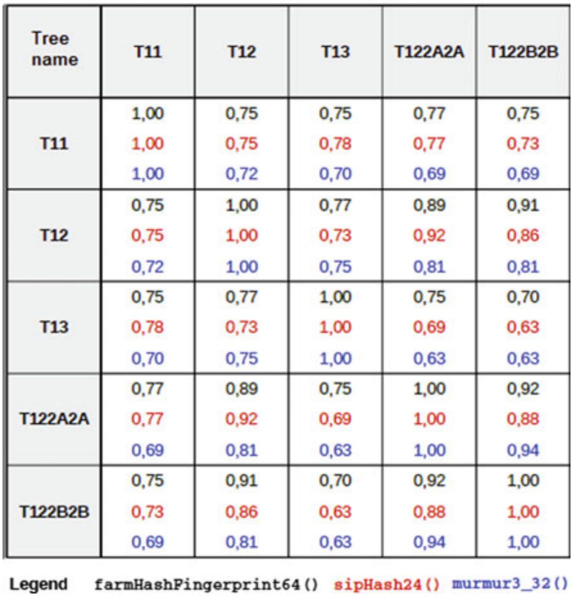
Results To present the results of the Simhash implementation, we choose 7 suitable types of root trees (they have different structure to each other and have different size), which you can see first 5 in Fig. 4. The calculated degrees of similarity of all pairs are arranged in a matrix of similarity as you can see in Fig. 4). Measured values of the degree of similarity are from the interval sim (A, B) $\in \langle 0.\,56 - 1.\,00 \rangle$.

The measured maximums correspond to similarities, not the minimum. The method is capable of detecting the equivalence of original structures. The method is sensitive enough and responds to minor changes to the original.

Experiment results have shown that the Simhash method can be used to compare root trees. The disadvantage remains high similarity values for very different

Fig. 4 Selected test trees





Tree name	T11	T12	T13	T122A2A	T122B2B
T11	1,00 1,00 1,00	0,75 0,75 0,72	0,75 0,78 0,70	0,77 0,77 0,69	0,75 0,73 0,69
T12	0,75 0,75 0,72	1,00 1,00 1,00	0,77 0,73 0,75	0,89 0,92 0,81	0,91 0,86 0,81
T13	0,75 0,78 0,70	0,77 0,73 0,75	1,00 1,00 1,00	0,75 0,69 0,63	0,70 0,63 0,63
T122A2A	0,77 0,77 0,69	0,89 0,92 0,81	0,75 0,69 0,63	1,00 1,00 1,00	0,92 0,88 0,94
T122B2B	0,75 0,73 0,69	0,91 0,86 0,81	0,70 0,63 0,63	0,92 0,88 0,94	1,00 1,00 1,00

Legend `farmHashFingerprint64()` `sipHash24()` `murmur3_32()`

originals. To fix this problem we could use hash functions with a longer fingerprint (especially cryptographic), other ways of original decomposition or possibly appropriate mathematical normalization of the calculated values. For usage in adaptive user interfaces, it needs to be set viable threshold. This threshold tells us if user structure includes into another cluster or not.

4 Conclusion and Future Work

This paper provided the background of methods for measuring the similarity of two sets. Then how we can use it with two structures like trees. It focuses on simhash algorithm. The main aim of this paper is to show novel view on how to use graph algorithms and clustering of trees into adaptive application structure.

In future work we plan to extend and improve the results achieved and also add other methods for calculating the degree of similarity of root trees. Then we want to focus on add this concept into adaptive structures of software [13, 15].

Acknowledgements Research described in the paper was supervised by doc. Ing. Karel Richta, CSc., FEE CTU in Prague and supported by the Czech Student Grant Agency under grant No. SGS18/185/OHK3/3T/13.

References

1. Bookstein A, Kulyukin VA, Raita T (2002) Generalized hamming distance. Inf Retr 5(4):353–375. https://doi.org/10.1023/A:1020499411651
2. Cerny J (2010) Zakladni grafove algoritmy. http://kam.mff.cuni.cz/~kuba/ka/ka.pdf
3. Charikar MS (2002) Similarity estimation techniques from rounding algorithms. In: Proceedings of the thiry-fourth annual ACM symposium on theory of computing, STOC '02. ACM, New York, NY, USA, pp 380–388. https://doi.org/10.1145/509907.509965
4. Chawathe SS (1999) Comparing hierarchical data in external memory. In: Proceedings of the 25th international conference on very large data bases VLDB '99. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 90–101. http://dl.acm.org/citation.cfm?id=645925.671669
5. Dehmer M, Mehler A (2007) A new method of measuring similarity for a special class of directed graphs. Tatra Mt Math Publ 36:1–22
6. Demel J (2002) Grafy a jejich aplikace, 1 edn. Academia, Praha (kveten)
7. Fr´echet MR (1906) Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Matematico di Palermo (1884--1940) 22(1): 1–72
8. Gary CHARTRAND, Ping ZHANG, L.L (2016) Graphs & digraphs, 6th edn . Chapman & Hal-l/CRC Press, Boca Raton
9. Jaccard P (1901) Etude de la distribution florale dans une portion des alpes et du jura 37:547–579. https://doi.org/10.5169/seals-266450
10. Katz J, Lindell Y (2014) Introduction to modern cryptography, 2 edn. Chapman & Hall, London
11. Rivest R (1992) The md5 message-digest algorithm. https://www.ietf.org/rfc/rfc1321.txt
12. Sebek J, Cerny T (2016) Aop-based approach for local data management in adaptive interfaces. In: 2016 6th international conference on IT convergence and security (ICITCS). pp 1–5
13. Sebek J, Cerny T, Richta K (2016) Adaptive application structure design for java ee applications. In: Proceedings of the international conference on research in adaptive and convergent systems. RACS '16, ACM, New York, NY, USA, pp 159–164. http://doi.acm.org/10.1145/2987386.2987417
14. Sebek J, Richta K (2016) Impact of users emotion on software adaptation1. Databases, Texts p. 1
15. Sebek J, Richta K (2016) Usage of aspect-oriented programming in adaptive application structure. New Trends Databases Inf Syst 217–222
16. Sedlacek J (1981) Uvod do teorie grafu, 3rd edn. Academia, Praha
17. Selkow SM (1977) The tree-to-tree editing problem. Inf Process Lett 6:184–186. https://doi.org/10.1016/0020-0190(77)90064-3
18. Tarjan R (1972) Depth-first search and linear graph algorithms. SIAM J Com- put 1(2):146–160. https://doi.org/10.1137/0201010
19. Tekli J, Chbeir R, Y´etongnon K (2007) Efficient xml structural similarity detection using sub-tree commonalities. In: SBBD. Brazil

20. Yang R, Kalnis P, Tung AKH (2005) Similarity evaluation on tree-structured data. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data, SIGMOD '05. ACM, New York, NY, USA, pp 754–765 http://doi.acm.org/10.1145/1066157.1066243
21. Zhen-kun W, Wei-zong Z, Ouyang-Jie, Peng-fei L, Yi-hua D, Meng Z, Jin-hua G (2010) A robust and discriminative image perceptual hash algorithm. In: 2010 fourth international

Part VIII
Web Technology

Nash Equilibrium Solution for Communication in Strategic Competition Adopted by Zigbee Network for Micro Grid

Seung-Mo Je and Jun-Ho Huh

Abstract In a complex Zigbee network for Micro Grid communication, a Zigbee node is affected by other Zigbee nodes and thus must predict their behaviors to communicate without collision. Therefore, this paper discusses the method of the network configuration according to Nash equilibrium for communication in strategic competition adopted by Zigbee network. It intended to show the efficiency increased by application of game theory through the simulation of communication under a competitive situation using OPNET. For the simulation, it configured OPNET for the worst situation of the Zigbee network communication environment, which were the distance between the nodes, the transmit power of node, and the number of nodes to reach avoid. It indicated that all nodes must know the information of when the communication started and ended or avoid the worst or avoid the situation to configure the Nash equilibrium under the strategic competition of Zigbee network.

Keywords Nash equilibrium · Solution · OPNET modeler · Riverbed modeler
Zigbee · Game theory · Smart grid · Micro grid

1 Introduction

There have recently been attempts at applying mathematical theories to a computer network. Game theory explains the existence of mutual dependency of strategic actors when an actor's utility or payoff is not limited to oneself but also by other actors' actions. Under such dependency, it is rational to decide one's action after predicting actions of others. Game theory is the science of analyzing the utility

S.-M. Je
Department of Computer Science Education, Korea University, Seoul, Republic of Korea
e-mail: jsm3316@korea.ac.kr

J.-H. Huh (✉)
Assistant Professor of Department of Software, Catholic University of Pusan,
Busan, Republic of Korea
e-mail: 72networks@cup.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_58

of decision making under such mutual dependent environment and predicting the result. If there is a strategic interaction in network configuration, we can consider each node of the network to be an actor and abstract it with game theory for the analysis. Game theory can be mainly divided into two broad categories. They are the simultaneous game and sequential game. A simultaneous game is like Rock Paper Scissors in which all players make decisions without knowledge of the actions of other players. A sequential game is like chess in which players obtain the information of other players' actions and then make a decision on subsequent action based on the information. One may consider a configuration of home networks using Zigbee as a sequential game since the node communications do not occur simultaneously. However, It is rational to consider it as a simultaneous game since the behavior of a node can only be known when the data transmitted by the node arrives, and not during the data are in transfer.

2 Related Works

The studies for both the Smart Grid and Micro Grid have been conducted rather actively in Europe and the game theory was applied in some of these studies. Since the electricity (power) market for the Smart Grid and Micro Grid systems are not fully established yet, the studies were carried out on the premise of the Pool Models using the game theory.

Saad et al. [1–3] organized the cases where the game theory has been applied to the Smart Grid and classified them into three categories. The first category includes the cases where the theory was applied to the Micro Grid distribution network. Here, the Micro Grid distribution network refers to a power distribution network where a multiple number of Micro Grids are being connected with some power companies, in a whole Micro Grid power network. This category deals with the issues in power exchanges between power companies or between each Grid in a free power market, assuming that such market exists. Saad et al. [4] analyzed the cooperative ties between the Micro Grids with the game theory. Friedman et al. [5] and Weaver and Krein [6] analyzed the relationships in the power exchange market where the double auctions had taken place. Using the game theory, Kasbekar and Sarkar [7] analyzed the pricing models in a competitive power exchange market where several independent micro Grids experience power shortage or surplus so that one has a chance to sell its surplus to the other. This study is similar to ours in terms of pricing but the difference is that our study focuses on the government-leading power management policy which is to encourage the Micro Grid-oriented power distribution network. The second category is for the demand management. This involves improvement of power production efficiency by controlling the power demands through differentiated pricing over time and/or adjustment of time periods for power uses. Mohsenian-Rad et al. [8] applied the non-cooperative game theory to solve the problem associated with scheduling of power supply plan in accordance with power use patterns. That is, by studying the timeslots of operations of consumers' household appliances, the supply plan can be rescheduled adequately. Assuming that each consumer in each household has the Smart Power Storage Equipment, Vytlingum et al. [9] also analyzed this category with the non-cooperative game theory. They analyzed when and how much consumers purchase power to store at their households or just use power directly without storing it. Their analysis also included the problem of how to control the power loads resulting from mass demands for the power storage by the majority of consumers. Finally, the game theory was applied to the issue associated with how to implement a network when focusing on the information delivery scheme. The Power Line Communication (PLC), WiFi and Mobile/Cable Communication Networks are currently available for network construction but Saad et al. [1–3] pointed out that the power line-oriented method has many technical problems while it's most promising in the short run (i.e., in costwise) as it has limited capacity and rapid signal attenuations in proportion to the communication distances. These problems are obstacles to mutual information exchanges in a network and there still remain many technical challenges. Authors propose the multi-hop network construction and its technique to overcome such problems. Saad et al. [1–3] tried to solve the problems in configuring the Smart Grid network using the non-cooperative game theory. Meanwhile, Gamma et al. [10] analyzed the benefits obtained from the cooperation between distributed power generators within the Micro Grid with the cooperative game theory by studying the costs and gains resulted from the cooperation, load and consumption curves and economical distribution/deployment of loads, for the three virtual diesel power generators. The study also involved technical aspects to construct a cooperative network.

Meanwhile, in the Republic of Korea, Hak-Jin Kim et al. [1] conducted a research in relation to the decision makings for all the regions' Micro Grids that experience power surplus or shortage focusing on the electricity prices by analyzing the models and deducting an appropriate solution [11].

3 Nash Equilibrium Solution for Communication in Strategic Competition Adopted by Zigbee Network for Micro Grid

As shown in Fig. 1, description of a game requires the players participating in the game, a set of strategies that each player can take, and the payoff of each strategy.

Although a home network applying Zigbee will typically have multiple nodes as the players $I = \{1, 2\ldots, N\}$, this paper limits the number of players $N = 2$ for convenience, clarity, and legibility. Also, a strategy refers to all actions that a player can take, and the set of strategies of a play can be denoted $Si = \{1, 2, 3, \ldots, K\}, \forall i \subseteq I$. In a Zigbee network, a node has two strategies of sending and not sending a signal.

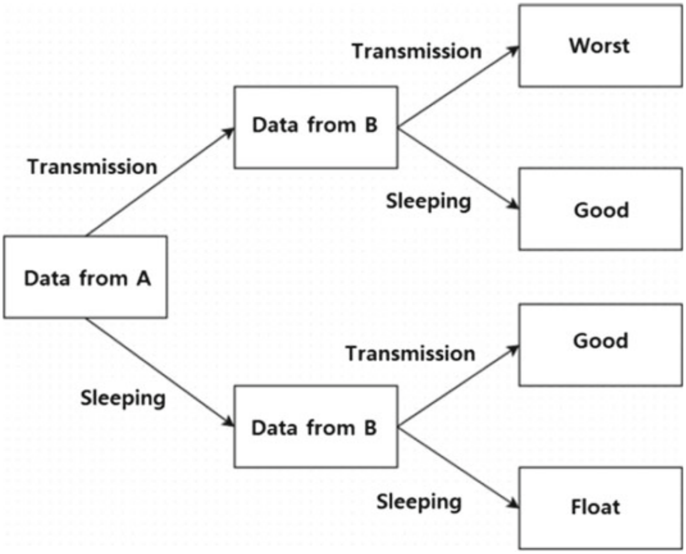
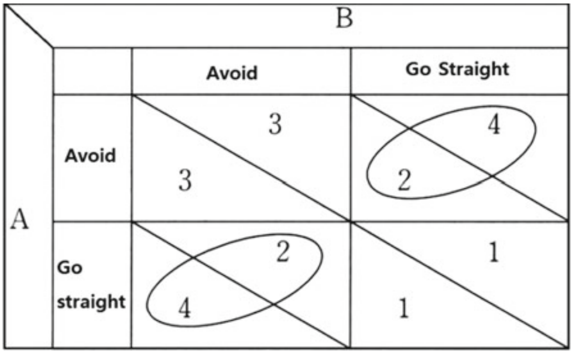


Fig. 1 Extensive form game theory of Zigbee communication



		B	
		Avoid	Go Straight
A	Avoid	3 3	4 2
	Go straight	2 4	1 1

Fig. 2 Normal form game theory of Zigbee and Nash equilibrium [12]

The strategy space is defined when the strategy sets of all players are defined, and it can be mathematically expressed as Π (i = 1 ~ N) Si. A strategy pair is the pair of strategies selected by all players, and the strategy pair of each player can be expressed as a vector of (1XN). It is expressed as the payoff matrix in the above normal form game. The payoff is a function of πi: S->R. The payoff is the quantified result of the strategy taken by a player. They are the values like those shown in Fig. 2 and can be expressed in a matrix.

The communication of two Zigbee nodes is the same as the dialog between two people. Let's assume that there are persons A and B. The person A is trying to deliver a speech to the person B while the person B is trying to deliver the speech b to the

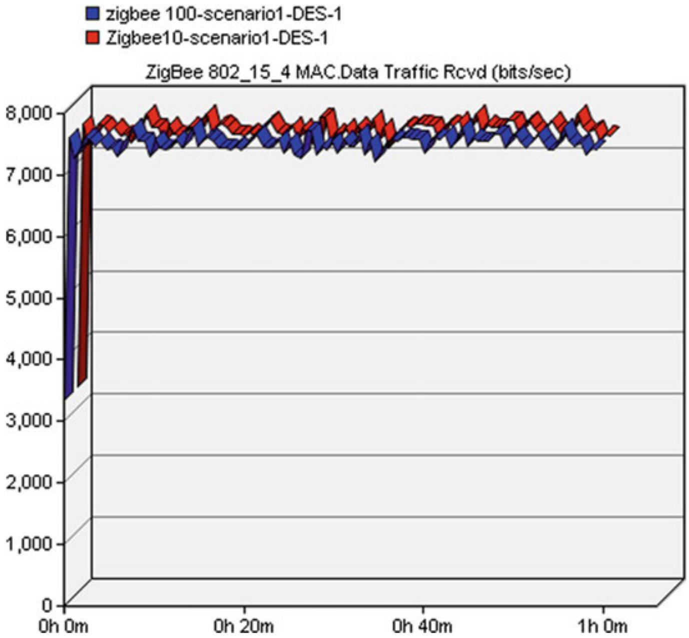


Fig. 3 Results of Zigbee 10 scenario and Zigbee 100 scenario

person A. However, the communication will not occur if both personal talk at the same time. It corresponds to two Go Straights and the payoff of 1.1 in a normal form game. It is equivalent to the worst situation in an extensive form game. There will be no communication also if neither person talks, waiting to hear what the other person says. It corresponds to two Avoids and the payoff of 3.3 in a normal form game. Although it can be considered to be better than the previous case since a third person can use the quiet environment, it is still not an ideal situation in communication aspect. It is ration to assume that there will be better communication if a person talks first and then the other person talks after the first person's talking is over. It is the Nash equilibrium solution in a normal form game.

Building such a Nash equilibrium network will require the information of when the other node will begin communication to avoid the worst or avoid situation. Or it can avoid the avoid situation by requesting the communication of a node in an avoid condition and assigning a priority to a node in the worst situation.

Therefore, this paper discusses the method of the network configuration according to Nash equilibrium as shown in Fig. 2 for communication in strategic competition adopted by Zigbee network. It intended to show the efficiency increased by application of game theory through the simulation of communication under a competitive situation using OPNET. Figure 3 shows the comparison of coordinator's traffic received when the Zigbee 10 scenario used the transmit power of 0.02 while

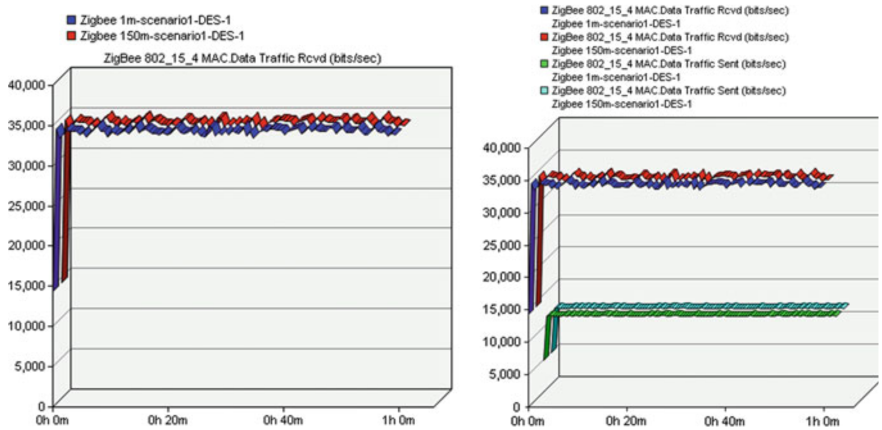


Fig. 4 Scenarios of 1 m distance and 150 m distance

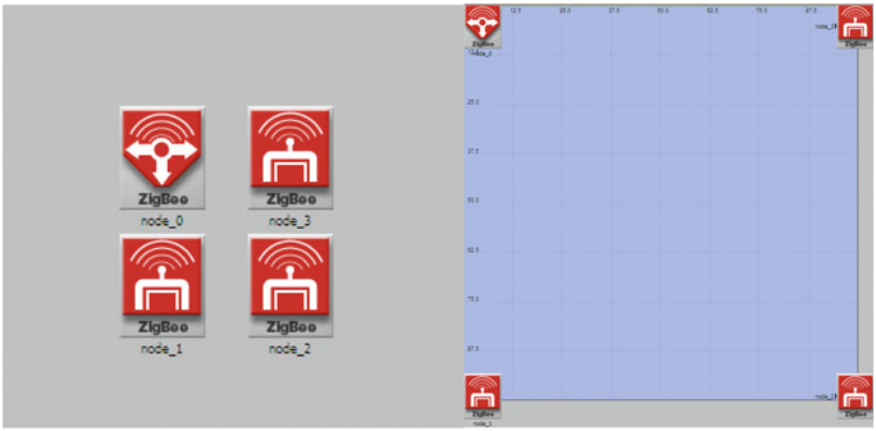


Fig. 5 Logical environment (Left) and office environment (Right)

the Zigbee 100 scenario used the transmit power at a distance of 150 m. The figure shows that there was no difference.

Figure 4 shows the comparison of Zigbee Send/Receive values at the distances of 1 and 150 m. The figure shows no difference between data Send and Receive values according to the distance. These results indicate that there is no visible change of transmitting efficiency according to the distance of 1–150 m between Zigbee nodes in OPNET and that the change of transmitting efficiency according to the transmit power of 0.02–10 was also very slight.

The difference between a logical environment and an office environment is that the logical configuration ignores the environmental factors such as distance while the office environment does as shown in Fig. 5.

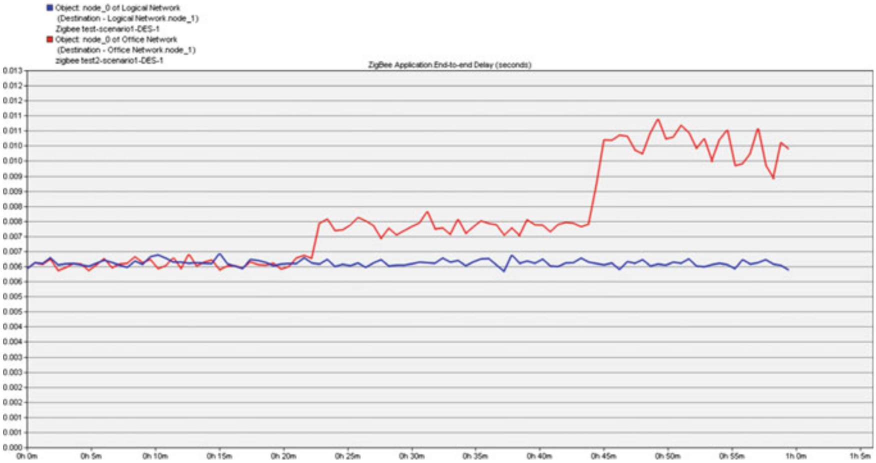


Fig. 6 ETE delays in the logical environment (Blue) and office environment (Red)

Figure 6 the only different setting between the logical environment and the office environment was the distance. It confirmed that the ETE delay increased as the distance increased (office environment). However, as the following graphs show, it is difficult to check the differences according to the Zigbee transmit power and distance. Also, Fig. 7 shows ETE delay according to Zigbee transmit power (0.2–10). Additionally, Fig. 8 ETE delay according to Zigbee distance (1–150 m). Therefore, the distance and transmit power are the consideration factors for the Zigbee simulation using OPNET Modeler 14.5 PL8. Also, Fig. 9 shows configuration of scenarios according to 12 nodes and 3 nodes.

The figures confirm that the ETE delays did not match. Although the differences were visible, the average ETE delay was still 0.027 s. It is due to the capacity of the Zigbee coordinator to facilitate the communication. The ETE delays of the two scenarios matched since the ETE delay was within the acceptable capacity. We doubled the number of nodes to 24 for the Zigbee network to observe the difference. Figure 10 shows comparison of ETE delay between 12 nodes and 3 nodes. Also, Fig. 11 shows Zigbee network with the number of nodes doubled to 24. Also, Fig. 12 shows the comparison of ETE delays of 24 nodes and 3 nodes. It indicates that the ETE delay of the scenario with 24 nodes was 0.005 s longer on average than the scenario with 3 delays.

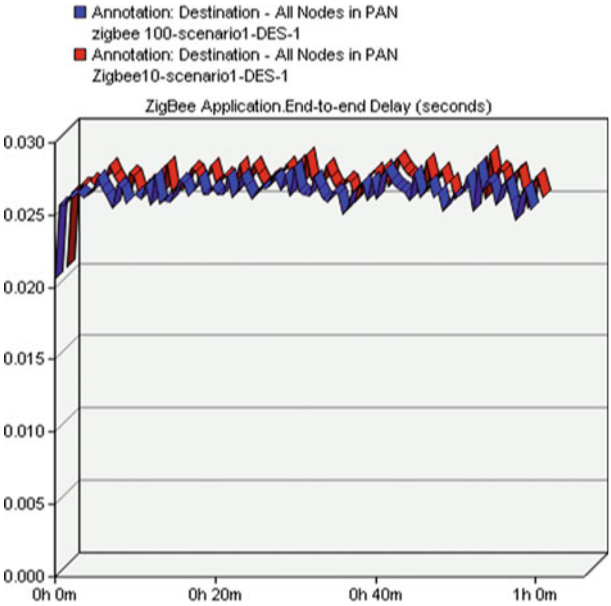


Fig. 7 ETE delay according to Zigbee transmit power (0.2–10) (Lift)

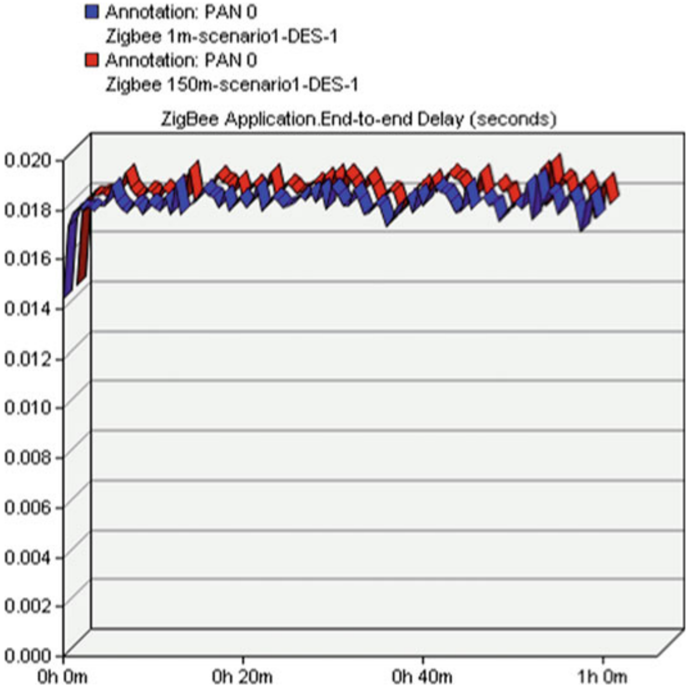
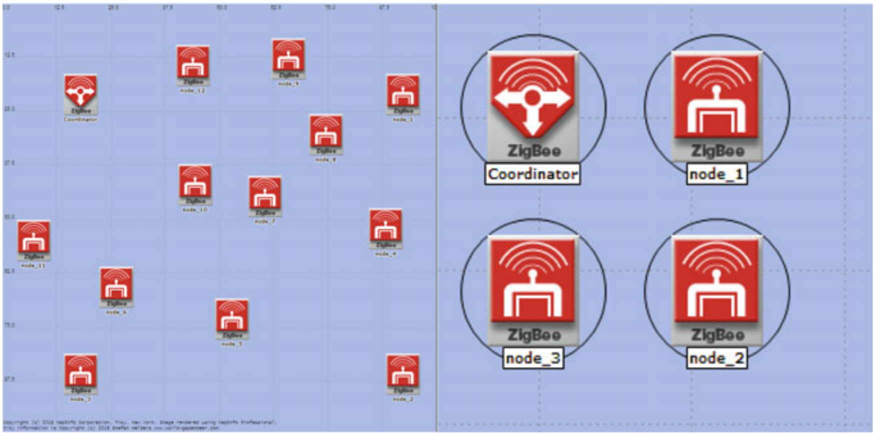


Fig. 8 ETE delay according to Zigbee distance (1–150 m) (Right)



(a) Configuration of scenario according to 12 Nodes

(b) 3 Nodes

Fig. 9 Configuration of scenarios according to 12 nodes and 3 nodes

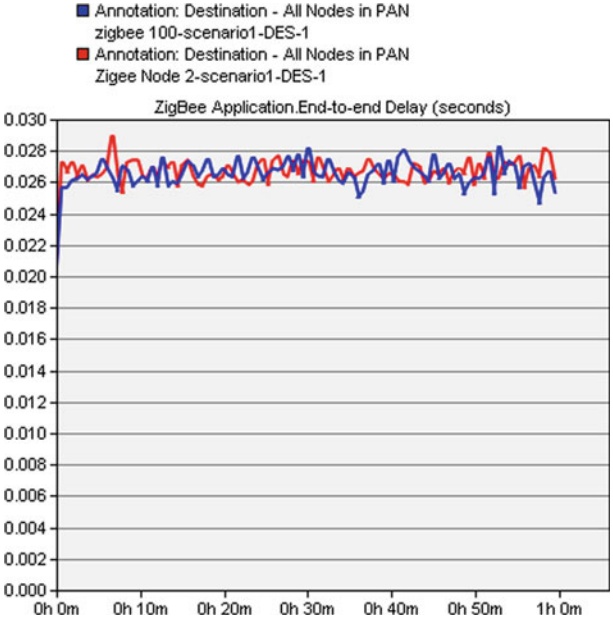


Fig. 10 Comparison of ETE delay between 12 nodes and 3 nodes

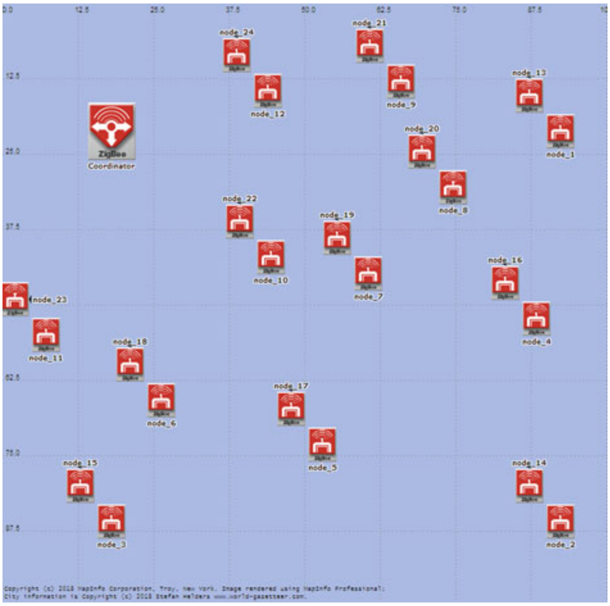


Fig. 11 Zigbee network with the number of nodes doubled to 24

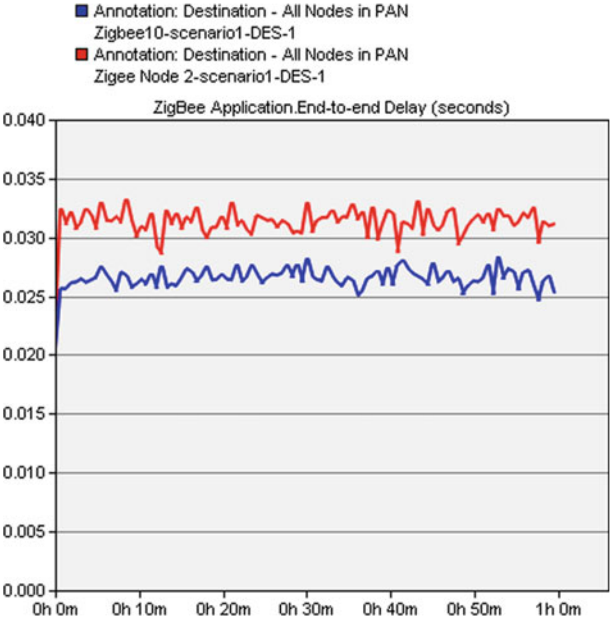


Fig. 12 Comparison of ETE delays of 24 nodes and 3 nodes

4 Conclusion and Future Works

The test indicates that all nodes must know the information of when the communication started and ended or avoid the worst or avoid situation to configure the Nash equilibrium under the strategic competition of Zigbee network for the Micro Grid. We varied three variables, the distance between nodes, the transmit power of node, and the number of nodes to search for the ways to avoid the worst and avoid situations and configure the Nash equilibrium under the strategic competition of Zigbee network in OPNET simulation. We did not observe any significant difference by varying the distance between the nodes and the transmit power but confirmed the visible change of the ETE delay according to the number of nodes. The future study intends to improve the efficiency in game theory aspect in assumed multiple nodes under the IoT environment by assigning priorities to the nodes and reducing the number of nodes available for communication by allowing the communication of the nodes with priority.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5077157).

References

1. Kim H-J, Lee SY (2015) A game theoretic decision making on the prices for surplus and deficiency of power generation in the micro grids. Korean Energy Econ Rev 14:1–34 (in Korean)
2. Saad W, Han Z, Poor HV, Başar T (2012) Game-theoretic methods for the smart grid: an overview of microgrid systems, demand-side management, and smart grid communications. Signal Process Mag 29(5):86–105
3. Saad W, Han Z, Poor HV (2012) A game theoretic approach for multi-hop power line communications. Game Theory Netw:546–561
4. Saad W, Han Z, Poor HV (2011) Coalitional game theory for cooperative micro-grid distribution networks. In: IEEE International conference, proceedings of in communications workshops (ICC), 5–9 June, Kyoto, Japan, pp 1–5
5. Friedman D, Friedman DP, Rust J (1993) The double auction market: institutions, theories, and evidence. Boulder, CO, Westview
6. Weaver WW, Krein DP (2009) Game-theoretic control of small-scale power systems. IEEE Trans Power Deliv 24(3):1560–1567
7. Kasbekar GS, Sarkar S (2012) Pricing games among interconnected microgrids. In: Proceedings of power and energy society general meeting, IEEE, 22–26 July, San Diego, USA, pp 1–8
8. Mohsenian-Rad H, Wong VWS, Jatskevich J, Schober R, Leon-Garcia A (2010) Autonomous demand side management on game-theoretic energy consumption scheduling for the future smart grids. IEEE Trans Smart Grid 1(3):320–331
9. Vytlingum P, Voice TD, Ramchurn SD, Rofers A, Gennings NR (2010) Agent-based microstorage management for the smart grid. In: Proceedings of international conference on autonomous agents and multiagent systems (AAMAS), Toronto, Canada, pp 39–46
10. Gamma PHRP, Gouvea MR, Torres GL (2007) Cost allocation by cooperation among distributed generators inside a micro grid using the cooperative game theory. In: Proceedings of 19th international conference on electricity distribution (CIRED'07), Vienna

11. Huh J-H, Seo K (2015) Hybrid advanced metering infrastructure design for micro grid using the game theory model. Int J Softw Eng Appl SERSC 9(9):257–268
12. Huh J-H, Je S-M, Seo K (2016) Design and configuration of avoidance technique for worst situation in zigbee communications using OPNET, Information Science and Applications (ICISA 2016). Lecture notes in electrical engineering, vol 376. Springer, Berlin, pp 331–336

Affordances for the Sharing of Domain-Specific Knowledge on Enterprise Social Media

L. G. Pee

Abstract Despite the general belief that enterprise social media (ESM) transcend traditional boundaries and offer access to knowledge in various domains, there has been a lack of understanding and empirical evidence for employees' willingness to expend the effort to share such knowledge when approached. This research-in-progress proposes that employees would be more willing to share domain-specific knowledge if they perceive the relevant ESM affordances for reducing the effort of codifying such knowledge. Preliminary results of a survey indicate that perceiving the affordance of visibility increases employees' willingness to share domain-specific knowledge, but the affordance of editability does not have a significant effect. Implications of the findings for research and practice are discussed.

Keywords Knowledge sharing · Social media · Affordance

1 Introduction

ESM allows employees to transcend rigid domain boundaries and easily identify others with a particular expertise. van Osch, Steinfield and Balogh [1] argue that ESM can assist in enhancing cross-boundary communication and decision making. In line with this, users of ESM identify the improved ability to find people with specific domain knowledge to be a key benefit of ESM [2]. Gibbs, Eisenberg, Rozaidi and Gryaznova [3] observed that employees use social media to traverse traditional barriers for cross-boundary knowledge sharing in practice.

But connecting employees in different domains is only part of the picture—the potential value of ESM's reach can only be realized if employees are willing to share domain-specific knowledge when approached. There has been a lack of understanding of employees' willingness to share domain-specific knowledge, despite the general belief that ESM is a rich source of such knowledge.

L. G. Pee (✉)
Nanyang Technological University, 31 Nanyang Link #05-06, Singapore 637718, Singapore
e-mail: peelg@ntu.edu.sg

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_59

Employees might not be willing to share domain-specific knowledge, as research on knowledge management suggests. Compared to general knowledge, sharing of domain-specific knowledge often requires the crossing of syntactic, semantic, and pragmatic boundaries [4] and therefore takes more effort. Knowledge codification effort has been found to reduce employees' willingness to share knowledge [e.g., 5]. This suggests that although ESM affords the possibility of accessing domain-specific knowledge, the effectiveness might be limited by employees' willingness to share such knowledge.

This study proposes that employees would be more willing to share domain-specific knowledge if they perceive the relevant ESM affordances for reducing the effort of codifying knowledge. Affordances are a user's perception of an object's utility, that is, possible actions linked to features [6]. Although an ESM's features are common to each person who encounters it, affordances are unique to the particular ways in which one perceives and uses the features. Research on ESM has identified several affordances. For instance, the affordance of *editability* makes it possible for users to revise knowledge shared by oneself or others over time to improve its quality [6].

Preliminary results of a survey measuring ESM users' perceptions and willingness to share domain-specific knowledge indicate that perceiving the affordance of visibility increases employees' willingness to share domain-specific knowledge, but the affordance of editability does not have a significant effect. The implications of these findings for research and practice are discussed in the concluding section.

2 Conceptual Background and Literature Review

2.1 Knowledge Specificity

Knowledge specificity is the extent to which knowledge supports a specific domain or function [7]. It is challenging to share domain-specific knowledge with outsiders because it requires crossing syntactic, semantic, and pragmatic boundaries [4]. Addressing syntactic boundary requires establishing a shared syntax or language for representing knowledge. Semantic boundary should recede when information that aids interpretation and understanding is provided. It is especially important to consider the individual, context-specific aspects of creating and sharing knowledge. To overcome the pragmatic boundary, it is important that the sharer is able to influence or transform the knowledge for application in other domains.

2.2 Enterprise Social Media Affordances for Knowledge Sharing

Research has shown that ESM are unique for knowledge sharing in that they “afford behaviors that were difficult or impossible to achieve in combination before” [6, p. 143]. Conventional knowledge sharing systems are often more centralized, formal, and reliant on users populating pre-constructed repositories, as compared to ESM that are more decentralized and allow emergent connections and continuous sharing [8].

Treem and Leonardi [6] identified four relatively consistent affordances enabled by ESM based on a review of preceding studies: editability, visibility, association, and persistence. Editability refers to the fact the individuals can craft and recraft a communicative act before it is viewed by others [6]. It allows individuals to modify or revise content they have already communicated, such as editing a typographical error or adding new information. These offer individuals the time to craft messages, and enable senders to compose messages to better convey the exact meaning intended. Editability also allows senders to consider the context in which their message is likely to be viewed and tailor it accordingly to improve its comprehensibility.

Visibility is the extent to which users can make their activities and knowledge that were once invisible or very hard to see visible to others in the organization [6]. From the knowledge providers’ perspective, visibility is closely tied to the presentation of self. The wide reach of social media can be used to show one’s expertise and competence. It is also useful for attracting the attention of specific audiences.

Associations are established connections between individuals (i.e., social ties) or between individuals and content [6]. Social media features affording association include list of friends and activities of related others. Many social media applications also have the capability of recommending new and potentially relevant associations (i.e., individuals and content) based on a user’s profile or activity.

Persistence refers to the degree to which knowledge shared on social media remains accessible in the same form as the original display after the sharer has finished his or her presentation [6]. Persistence mainly benefits knowledge seekers by helping to sustain knowledge over time even when sharers are not active, allowing knowledge reuse and thereby creating robust knowledge, and facilitating the accumulation and growth of knowledge. Since this study’s focus is willingness to share knowledge by potential providers rather than seekers, we do not consider persistence further.

3 Hypothesis Development

As discussed before, sharing domain-specific knowledge requires effort in crossing syntactic, semantic, and pragmatic boundaries [4]. The effort should reduce employees’ willingness to share such knowledge, as indicated by prior studies. He and Wei [5] observed that employees’ belief about knowledge codification effort significantly

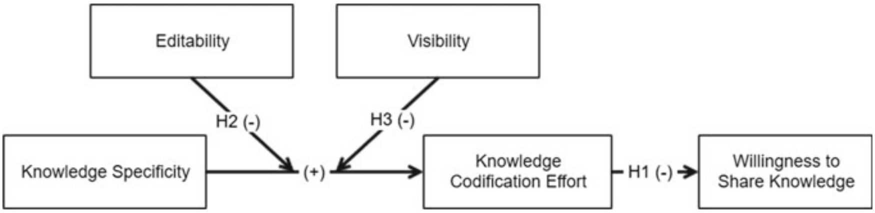


Fig. 1 Model of hypotheses

reduces their willingness to share knowledge using a KM system; A study on enterprise Wikis showed that codification effort hinders employees' knowledge sharing in the forms of article creation and editing [9]. Therefore, we hypothesize that:

H1: Employees are less willing to share knowledge that requires more codification effort (i.e., knowledge codification effort and willingness to share knowledge are negatively related).

The ESM affordance of editability could attenuate the perceived effort of sharing domain-specific knowledge. Editability allows one to split the effortful task of providing information about a domain, depicting the specific situations in which the knowledge is applicable, defining jargons, or even "translating" the knowledge for application in another context [10] into smaller, more manageable subtasks. High level of editorial control also allows sharers to gradually tailor their messages to the intended audience to improve its applicability to other domains or contexts. Accordingly, this study hypothesizes that (see Fig. 1):

H2: Employees are more willing to share domain-specific knowledge when they perceive the affordance of editability, because the affordance reduces the perceive codification effort of sharing such knowledge (i.e., editability negatively moderates the positive relationship between knowledge specificity and knowledge codification effort).

The affordance of visibility should also reduce the perceived effort of sharing domain-specific knowledge. Domain-specific knowledge is developed for a particular context or functional unit and tends to be invisible by outsiders [11]. Visibility makes it possible for employees to show that they are experts in a specific domain with deep and intimate understanding of the context [6]. Fulk and Yuan [12] argued that ESM provide reputational benefits and thereby help to address challenges related to motivation to share knowledge; Rode [13] found that the expectation for improvement in reputation significantly increased employees' willingness to share knowledge using ESM.

H3: Employees are more willing to share domain-specific knowledge when they perceive the affordance of visibility, because the affordance reduces the perceive codification effort of sharing such knowledge (i.e., visibility negatively moderates the positive relationship between knowledge specificity and knowledge codification effort).

Persistence is an affordance for knowledge seekers and is therefore not considered in this study of knowledge sharing, as explained before. As for the affordance of association, there is a lack of theoretical rationale for expecting that it addresses the challenges of sharing domain-specific knowledge, because the source of such knowledge tends to be confined to a specific domain or function and the broad associations or connections afforded by ESM are not likely to be especially useful. Nevertheless, the effect of association effects is controlled for in data analysis to better discern the relationships hypothesized.

4 Research Method

Data were collected through a survey since this study seeks to understand employees' willingness to share domain-specific knowledge based on their perceptions of ESM affordances. Data were collected from employees in two organizations. One was a large producer of specialty chemicals (e.g., coatings, additives, inorganic materials, performance polymers). The organization used an ESM built in house. The other organization was a public police force in Asia with about 38,000 employees, and engages in activities such as criminal investigation, police intelligence, and traffic policing. It used Workplace by Facebook as its ESM. Both chemical manufacturing and law enforcement organizations engage in highly knowledge-intensive activities and rely on knowledge and experience to perform effectively. The effect of organization was controlled for in data analysis to rule out the effect of organizational differences.

5 Data Analysis and Preliminary Findings

A total of 199 responses have been collected thus far and they were analyzed using Partial Least Squares (PLS), since the phenomenon investigated is relatively new and there is a lack of well-established measurement models [14]. The PLS approach is adequate for causal modeling whose purpose is prediction and theory building [14], which is in line with our research objective of understanding employees' willingness to share domain-specific knowledge on ESM.

Results of the preliminary analysis indicate that the measurement model developed had satisfactory reliability, convergent validity, and discriminant validity. Analysis of the structural model shows that knowledge codification effort reduces employees' willingness to share knowledge ($\beta = -0.17$, $p < 0.01$; H1 was supported). The codification effort is significantly attenuated by the affordance of visibility ($\beta = -0.13$, $p < 0.01$; H3 was supported) but not editability ($\beta = 0.03$, $p > 0.05$; H2 was not supported). Detailed results will be presented at the conference.

6 Discussion and Conclusion

This study's preliminary findings indicate that the sharing of domain-specific knowledge on ESM takes more effort and employees are therefore less willing to share, but those who perceive the affordance of visibility are more willing to do so. In contrast, the affordance of editability does not matter significantly for the sharing of domain-specific knowledge. For research, these findings deepen our understanding of the use of ESM for knowledge sharing by highlighting the importance of clarifying the relationships between knowledge attributes and ESM affordances. This study goes beyond a general notion of knowledge to account for the knowledge attribute of specificity and identifies the relevant affordance for addressing the challenges of sharing domain-specific knowledge. A new line of inquiry indicated by the findings is identifying the relevant affordances for other types of knowledge that tends to be sought using ESM, such as complex knowledge or emerging/volatile knowledge. This is also in line with the concept of affordance, which by definition needs to be understood relative to the use of a feature rather than the feature or the user per se.

For practice, the findings suggest that raising employees' awareness of the affordance of visibility increases their willingness to share domain-specific knowledge. For example, ESM affordances can be circulated in the form of usage tips; success stories involving the sharing of domain-specific knowledge on ESM could also be publicized.

References

1. van Osch W, Steinfield CW, Balogh BA (2015) Enterprise Social Media: Challenges and Opportunities for Organizational Communication and Collaboration City
2. Evans RD, Ahumada-Tello E, Zammit J (2017) Yammer: Investigating its impact on employee knowledge sharing during Product Development
3. Gibbs JL, Eisenberg J, Rozaidi NA, Gryaznova A (2014) The "Megapozitiv" role of enterprise social media in enabling cross-boundary communication in a distributed Russian Organization. Am Behav Sci 59(1):75–102
4. Carlile PR (2004) Transferring, translating, and transforming: an integrative framework for managing knowledge across boundaries. Organ Sci 15(5):555–568
5. He W, Wei K-K (2009) What drives continued knowledge sharing? An investigation of knowledge- contribution and seeking beliefs. Decis Support Syst 46(4):826–838
6. Treem JW, Leonardi PM (2013) Social media use in organizations: exploring the affordances of visibility, editability, persistence, and association. Ann Int Commun Assoc 36(1):143–189
7. Earl M (2001) Knowledge management strategies: toward a taxonomy. J Manag Inf Syst 18(1):215–233
8. Majchrzak A, Faraj S, Kane GC, Azad B (2013) The contradictory influence of social media affordances on online communal knowledge sharing. J Comput-Mediat Commun 19(1):38–55
9. Beck R, Rai A, Fischbach K, Keil M (2015) Untangling knowledge creation and knowledge integration in enterprise wikis. Journal of Business Economics 85(4):389–420
10. Hacker J (2017) Enterprise social networks: platforms for enabling and understanding knowledge work. Springer International Publishing, City
11. Subramani M (2004) How do suppliers benefit from information technology use in supply chain relationships? MIS Q 28(1):45–73

12. Fulk J, Yuan YC (2013) Location, motivation, and social capitalization via enterprise social networking. J Comput-Mediat Commun 19(1):20–37
13. Rode H (2016) To share or not to share: the effects of extrinsic and intrinsic motivations on knowledge-sharing in enterprise social media platforms. J Inf Technol 31(2):152–165
14. Henseler J, Ringle CM, Sinkovics RR (2009) The use of partial least squares path modeling in international marketing. Adv Int Mark 20:277–319

Part IX
Internet of Things

Development of an IoT-Based Construction Site Safety Management System

Seung Ho Kim, Han Guk Ryu and Chang Soon Kang

Abstract An effective construction safety management system is required for reducing damage caused by construction site accidents. However, construction site safety management systems are mainly operated only at large scale, so there is a lack of safety management system that can be operated at a low cost even in smaller construction sites. In this paper, we propose an Internet of Things (IoT)-based construction site safety management system which can be operated at low cost not only at large construction sites but also at smaller construction sites. A prototype for the proposed system has been developed using a beacon technology, smartphone application, and sensors, Zigbee, WiFi, and LTE to monitor field workers and outsiders approaching the hazard zones at all times. The prototype system also provides danger alarm to safety managers in construction sites and at remote sites. It is expected that the developed system can effectively prevent safety accidents in large-scale and small-scale construction sites.

Keywords IoT/M2M · LTE · Zigbee · WiFi · Sensor · Construction site
Safety management · Web server · Smartphone application

S. H. Kim (✉)
Department of Eco-friendly Offshore Plant FEED Engineering,
Changwon National University(CNU), Changwon, Republic of Korea
e-mail: shk0529@changwon.ac.kr

H. G. Ryu
Department of Architecture Engineering, CNU, Changwon, Republic of Korea
e-mail: hgryu@changwon.ac.kr

C. S. Kang (✉)
Department of Information & Communication Engineering, CNU, Changwon,
Republic of Korea
e-mail: cskang@changwon.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_60

1 Introduction

The Internet of Things (IoT) is a technology in which intelligent objects or devices such as sensors, communication devices, and computing devices are interconnected through wired and wireless communication networks to collect, share, and utilize information without human intervention [1].

According to the report of Korea Occupational Safety and Health Agency (KOSHA) on industrial accidents in 2016, 29% of the total industrial accidents were suffered by 26,570 workers, and 499 deaths were recorded due to industrial accidents [2]. In order to prevent safety accidents at the construction sites, the research combining construction safety systems and information communication technologies have been actively carried out. These studies include the safety systems which manages the condition of the workers and the construction site by attaching various sensors and cameras to the worker's helmets [3], sensor location monitoring and risk detection system for underground construction workers [4], location tracking of construction workers using MEMS sensors [5], construction site safety management and maintenance system utilizing augmented and virtual reality (VR/AR) technology [6], and building information modeling (BIM) that uses big data for design and construction at construction sites [7].

Even though the construction site safety management system is an urgent situation for the safety of the workers, it is difficult to operate an effective safety management system due to the operation cost that is lacking in small construction sites. In this paper, we propose an IoT-based construction site safety management system that can effectively prevent safety accidents of field workers in both large-scale and small-scale construction sites. The proposed system has been developed with a prototype using sensors, smartphone, management server, and short and wide-range wireless communication systems such as Zigbee, Wireless Fidelity (WiFi), and Long Term Evolution (LTE), etc.

2 System Design

In order to design the proposed safety management system we consider such requirements as shown in Table 1. The proposed system can setup, suspend, and release the hazard zones and provide the access information of field workers and outsiders through a manager's smart application.

The proposed system is designed considering the system requirements, in which the system consists of IoT-cones, worker safety check devices, the mobile gateway, the safety management server, and the safety manager application, as shown in Fig. 1.

An IoT-cone periodically transmits a beacon signal with a Zigbee-based beacon transmitter attached to detect workers approaching a hazard zone. In addition, a construction site safety manager receives a mode change message when changing the operation mode of the IoT-cone (i.e., setup/suspend/release). A worker safety check

Table 1 System requirements

Requirements	Details
Providing means of safety management of hazard zones	• IoT-cone can setup, suspend, release hazard zones of each floor
Providing real time risk information	• Safety manager application provides the floor image of each floor and hazard zones of the construction site together
Providing on-site alarms for fieldworkers or outsiders approaching hazard zones	• Zigbee and ultrasonic sensors provide on-site alarms when workers or outsider approach hazard zones
Providing alarms to the safety manager when field workers are approaching hazard zones	• Provide the safety manager with the names of field workers and locations while approaching the hazard zones

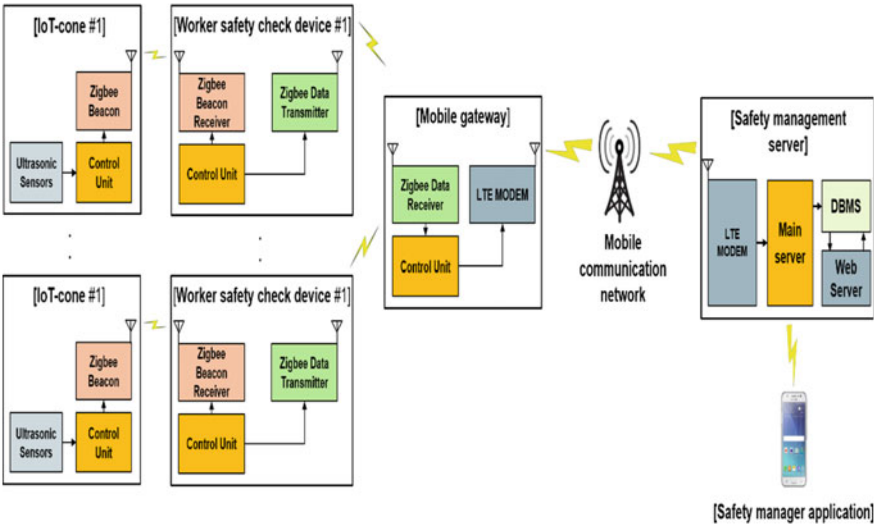


Fig. 1 Detailed configuration of the proposed system

device receives a beacon signal (RSSI: Received Signal Strength Indicator) from the IoT-cone and converts it to the distance between the transmitter (IoT-cone) and the receiver (approaching worker). When the worker approaches within a predefined threshold distance of a setup hazard zone, a built-in alarm device in the worker safety check device generates an alarm signal and informs the worker that the hazard zone is nearby. In addition, the worker safety check device transmits both IDs of the worker and the IoT-cone to a safety manager.

The mobile gateway acts as a data relay between an IoT-cone or a worker safety check device and the safety management server. The safety management server transmits the hazard zone-access notification message of the received worker or

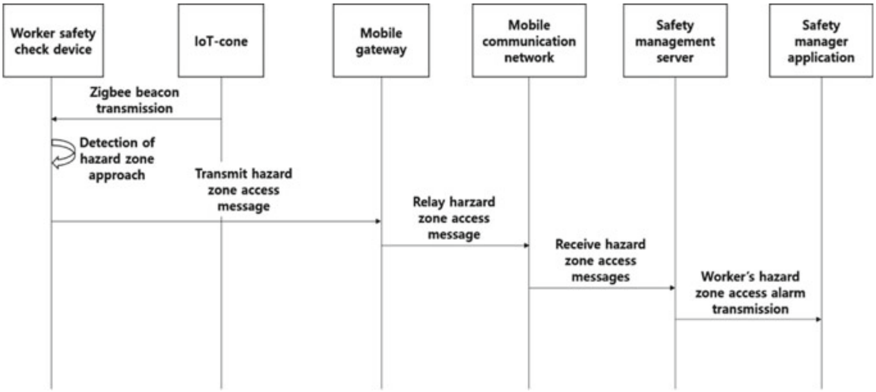

Worker safety check device
IoT-cone
Mobile gateway
Mobile communication network
Safety management server
Safety manager application
Zigbee beacon transmission
Detection of hazard zone approach
Transmit hazard zone access message
Relay harzard zone access message
Receive hazard zone access messages
Worker's hazard zone access alarm transmission

Fig. 2 Sequence diagram of the system

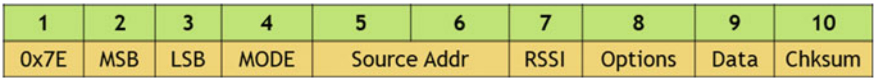
the external person (outsider) to a safety manager application. The safety manager application can send the setup/suspend/release command of an IoT-cone and confirm the location of the IoT-cone on the image of each floor. Additionally, when workers or outside persons access to the predetermined hazard zone, it receives the hazard zone access notification message transmitted from the safety management server and confirms the name and the location of the hazard zone. The sequence flow of the system is as shown in Fig. 2.

3 System Implementation

An IoT-cone connects the safety management server through the relay of mobile gateway and receives an operation mode change message upon the setup/suspend/release command from a safety manager. Also, the IoT-cone receives the measured values from an ultrasonic distance measurement sensor by using the power control of beacon transmission unit of Zigbee and an ultrasonic sensor and UART (Universal Asynchronous Receive and Transmit) communication. The beacon transmitter generates a beacon signal including the ID of an IoT-cone at a period of 100 ms. If the operation mode of the IoT-cone is the (in) suspend or release state, the beacon signal is not transmitted. The detection coverage of ultrasonic sensor of outsiders is at 15 cm ~ 6 m. The detection of outsiders are done in every 300 ms cycle, and an alarm device (LED strip, buzzer) is activated if an outsider is detected, after that an alarm message is sent to the mobile gateway to send a hazard zone-access notification message.

A worker safety check device extracts a RSSI value and an ID of an IoT-cone among the frames of Zigbee beacon signal received from an IoT-cone. The frame format of a Zigbee beacon signal received from the IoT-cone is shown in Fig. 3.

The RSSI measurement value is converted to the distance between an IoT-cone and a worker, and is stored together with an ID of the IoT-cone. When the conversion



1	2	3	4	5	6	7	8	9	10
0x7E	MSB	LSB	MODE	Source Addr		RSSI	Options	Data	Chksum

Fig. 3 Beacon receive frame format

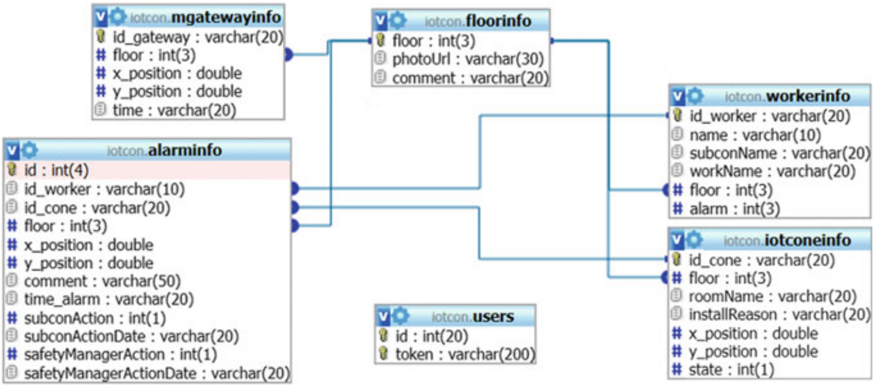


Fig. 4 Safety management database ER (Entity-Relationship) diagram

distance is less than 0.5 m to a hazard zone, the IoT-cone calculates the average distance value using the previous measured values of 5 times, and if the average distance is similar to the conversion distance, then the alarm device of the worker is activated. When the worker approaches the hazard area, the worker's hazard zone-access notification message is transmitted to a mobile gateway through the Zigbee transmitter of the safety check device.

A mobile gateway receives a hazard zone-approach detection notification message through Zigbee (and WiFi) between an IoT-cone and a worker safety check device and transmits the received notification message to the safety management server through an LTE modem. The safety management server receives the hazard zone-access notification message and a command of hazard zone-setup/suspend/release and stores both the message and command to database (DB) of the safety management server. The DB tables of the safety management server consists of information of mobile gateway, floor information, worker information, IoT-cone information, hazard zone access alarm list, and manager information for smart phone alarm. The relationship between the DB tables is shown in the Fig. 4. It then transmits the hazard zone- access notification message to the safety manager application through the web server to a safety manager application.

A safety manager application is implemented with the Android OS, in which the application can setup, suspend, and release hazard zones, and provide danger alarms for field workers or outsiders. Figure 5 shows the developed IoT-cone and worker safety check device.

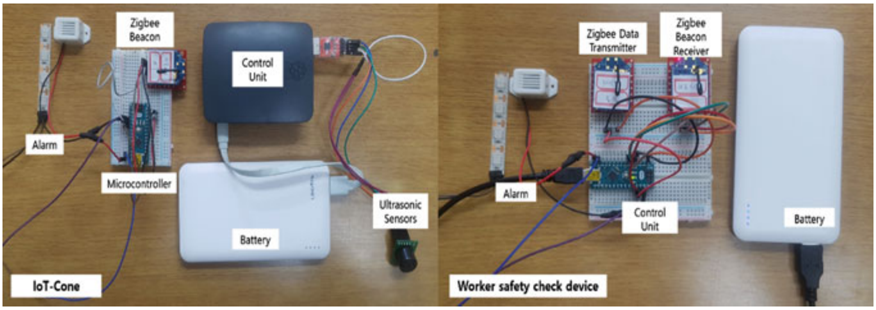


Fig. 5 Developed construction site safety management system

4 Field Trials and Discussions

At a construction site, a safety manager can select the ID of an IoT-cone and the location of a hazard zone from the smartphone application to set up the IoT-cone. The hazard zone setting information is stored to the DB of the safety management server. When the setup of the hazard zone is normally completed, the safety management server transmits the setup completion information to the safety manager application. An example screen for setting up the IoT-cone is shown in Fig. 6.

In order to detect workers or outsiders approaching a hazard zone, a worker safety check device receives the Zigbee beacon signal value (RSSI in *dBm* unit) from the IoT-cone and converts it into a distance (meter) between a worker and the hazard zone. The conversion distance according to the RSSI value is shown in Fig. 7. As a result, the conversion distance of the Bluetooth beacon signal [8] and the conversion distance of the ZigBee beacon signal are similar to each other.

When a worker approaches a hazard zone within 3 m, the safety check device of the worker turns on a warning light. On the other hand, the worker approaches

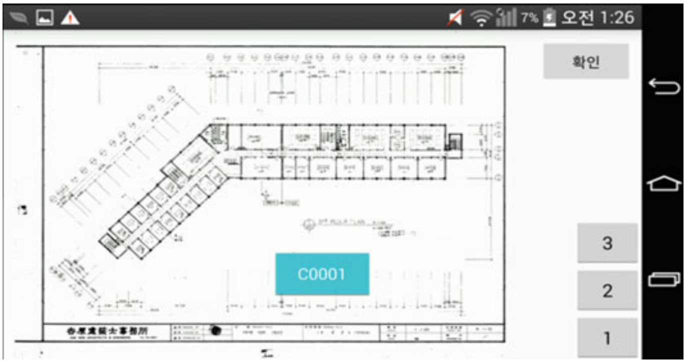


Fig. 6 Screen for hazard zone setting

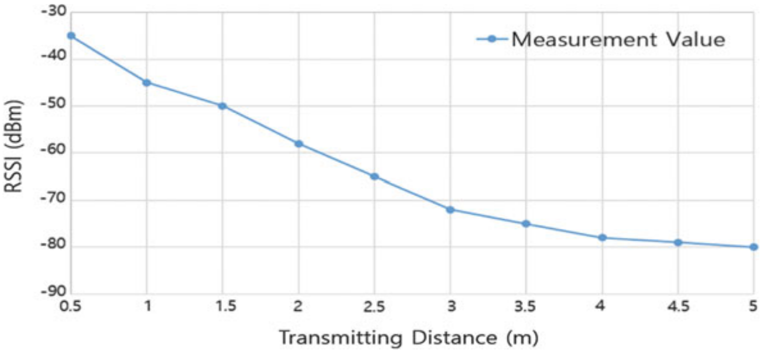
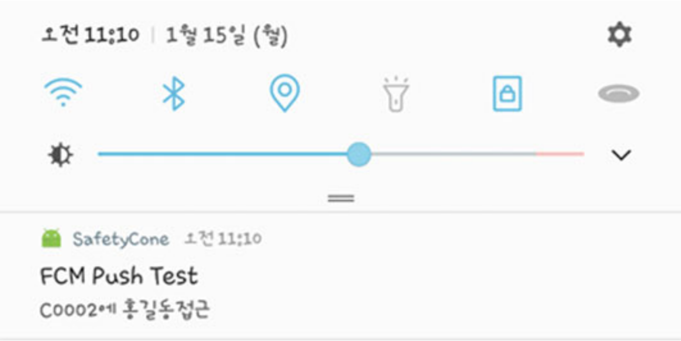


Fig. 7 The conversion distance according to the measured RSSI value


오전 11:10 | 1월 15일 (월)

SafetyCone 오전 11:10

FCM Push Test

C0002에 홍길동접근

Fig. 8 Worker's hazard zone-access displayed in the application

within 0.5 m, the worker safety check device generates a warning sound together with a warning light. The safety management server receives a hazard zone-access notification message and stores it to the DB. After receiving the notification message, the safety manager can recognize the location of the worker and the hazard zone as shown in Fig. 8.

5 Conclusion

In this paper, we have proposed an IoT-based construction site safety management system which can be operated at low cost in both small and large-scale construction sites. The development system consists of IoT-cones, the worker safety check devices, the mobile gateways, the safety management server, and the smartphone application of the safety manager. The IoT-cones installed in hazard zones detect the approach of workers and outsiders, generate warning signals in the hazard zones when approaching the hazard zones, and provide the situation to the construction safety manager. The proposed system has been developed with a prototype using ultrasonic sensors, embedded systems, such wireless communication systems as the Zigbee, Wireless Fidelity (WiFi), and Long Term Evolution (LTE). It is expected that the developed prototype can drastically reduce the number of accidents when applied to the construction sites. Furthermore, it can be used to build big data for preventing safety accidents at the construction site by using the IDs of IoT-cones and workers that access to the hazard zones.

Acknowledgements This work was supported by two grants; the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1A2B4003722) and the Brain Korea 21 plus projects of the NRF funded by the Ministry of Education of Korea.

References

1. ETSI (European Telecommunication Standard Institute), M2M Standards: how to enable the internet of the Future, M2M workshop (2010)
2. Korea Occupational Safety & Health Agency, Industrial Disaster Analysis 2016 (2017)
3. Lee CH, Kim KH, Kim JW, Choi SB (2017) Construction site safety management system using zigbee communication. J Inst. Electron Inf Eng 54(3):39–51
4. Zhou C, Ding LY (2017) Safety barrier warning system for underground construction sites using internet-of-things technologies. Autom. Constr. 83:372–389
5. Kim JY, Ahn SS, Kang JH (2012) Development of location/safety tracking system for construction site workers by using MEMS sensors. J Inst Electron Inf Eng 49(1):12–17
6. Alama MF, Katsikasb S, Beltramelloc O, Hadjiefthymiadesa S (2017) Augmented and virtual reality based monitoring and safety system: a prototype IoT platform. J Netw Comput Appl 89:109–119
7. Riaz Z, Parn EA, Edwards DJ, Arslan M, Shen C, Pena-Mora F (2017) BIM and sensor-based data management system for construction safety monitoring. J Eng Design Technol 15(4):738–753
8. Mohamed E, Liu F, Jadi Y (2015) Indoor location position based on bluetooth signal strength. In: Information science and control engineering (ICISCE), 2015 2nd international conference, pp 769–773

Internet of Things: Current Challenges in the Quality Assurance and Testing Methods

Miroslav Bures, Tomas Cerny and Bestoun S. Ahmed

Abstract Contemporary development of the Internet of Things (IoT) technology brings a number of challenges in the Quality Assurance area. Current issues related to security, user's privacy, the reliability of the service, interoperability, and integration are discussed. All these create a demand for specific Quality Assurance methodology for the IoT solutions. In the paper, we present the state of the art of this domain and we discuss particular areas of system testing discipline, which is not covered by related work sufficiently so far. This analysis is supported by results of a recent survey we performed among ten IoT solutions providers, covering various areas of IoT applications.

Keywords Internet of things · Quality assurance · Testing methodology · Test strategy · Integration testing · Security · Interoperability · Integration issues

1 Introduction

In last two decades, the Internet of Things (IoT) solutions started to emerge from the initial pioneering visions to regular industrial solutions, which are present in our everyday lives. The lively development of these solutions brings also a number of challenges [1, 2]; as common examples, we can discuss the insufficient level of standardization, legislation and quality assurance techniques, as well as security and privacy concerns [3–6]. In this paper, we focus on the quality assurance and testing techniques for the IoT domain. Despite the fact, that some of the areas are intensely covered by the literature (security and privacy are the typical representatives), in the area of systematical testing and quality assurance methodologies, much less work exists. In this paper, we present an overview of the domain and identify the areas,

M. Bures (✉) · B. S. Ahmed
FEE, CTU in Prague, Karlovo nam. 13, 121 35 Praha 2, Czech Republic
e-mail: miroslav.bures@fel.cvut.cz

T. Cerny
Computer Science, Baylor University, Waco, TX, USA

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_61

Table 1 Consequences of IoT issues for testing methods

Issues	Consequence for testing methods
1, 5, 9	Demand for comprehensive method to define efficient test strategy for IoT solutions
2, 3, 4, 6, 7	Increased demand for security testing, including privacy aspects
3, 8	Demand for more efficient methods how to select economic but representative platform variants to test
3, 5, 8	Increased demand for more efficient integration testing, if possible, automated
1, 3, 5	Test automation in general, as the number of variants seems not feasible to be tested manually
9	Testing of behavior of the IoT solutions under limited connection and various edge conditions is needed, especially for life-critical systems

which we consider relevant for the further research. This analysis is supported by discussion of the specifics of IoT solutions having an impact on particular testing techniques and methods, together with a literature survey and with a survey among ten IoT solutions providers, which provided us with different, independent viewpoints on the problem discipline.

The paper is organized as follows. Section 2 analyzes principal issues of IoT solutions, leading to challenges in IoT quality assurance. Section 3 summarizes the state of the art in this domain. Section 4 presents the results of the recent survey among IoT solutions providers. In Sect. 5 we discuss the results and we identify the quality assurance areas, which have to be covered by a more intense research. The last section concludes the paper.

2 Principal IoT Issues with Impact on Testing Techniques

A number of discussions have been conducted regarding the IoT issues, for instance in [1–5, 7, 8]; however, during our literature survey, we have not found a systematic analysis, identifying what is the impact of these specifics to particular software testing methods and techniques. Hence, we provide such an analysis in this paper. In the following section, we identify several typical issues of IoT solutions and we number them by IDs. Next, in Table 1, we map these issues with direct consequences they have on the testing and quality assurance process.

Issue 1. From the business and economic viewpoint, competition in IoT business is having a direct impact on the conditions, in which these solutions are developed. This competition triggers a demand to lower prices of the manufactured IoT devices, as well as it creates a pressure to shorten time to market.

Issue 2. In specific applications of the IoT as the sensor networks or camera networks are, the devices can be located in places, which makes them easily accessible by an attacker; on the other hand, difficult to check by the service provider periodically. These devices can act as a vulnerable point to the entire network.

Issue 3. Another related issue is a low possibility to update certain types IoT devices. Either due to low production costs or energy consumption issues it is not possible to update some types of devices, which is typical for sensor networks. This has two consequences: (1) known security defects can be exploited by a potential attacker, and (2) inability to update the device firmware can lead to significant number of various versions of the devices used in production run of the service; these variants need to be tested, which increases the costs of the testbed and also number of variants to test.

Issue 4. The IoT devices powered by battery or solar energy lead engineers to minimize the power consumption of the device. This can lead to the implementation of lightweight authorization and security algorithms, exposing these IoT devices as a weak entry point to the whole network.

Issue 5. Compared to common web-based internet solutions, testing IoT solutions is specific from another viewpoint. When testing the web-based systems, we usually assume, that the lower physical layers (hardware, network protocols, operational systems, application servers etc.) are tested sufficiently already by supplier parties. Hence, we focus the system testing effort mainly on the application and integration levels. In IoT, the situation is utterly different. Compared to web-based solutions, there is a much more extensive variety of used standardized protocols [9]. Moreover, a number of proprietary protocols are used in the current IoT solutions. Thus, testing IoT services usually involves specific testing of the lower layers of the system; when a service involves development of the own IoT devices, we need to test also this hardware.

Issue 6. IoT devices are connected to the Internet network, which has at least two consequences: (1) number of links between connected devices will grow rapidly, and (2) weakly secured device can act as an entry point to the entire network.

Issue 7. In a number of IoT devices, the user can have low insight into the internal mechanism of a device; also, if a device is updated, the user can have low control about these updates. Combined with GPS, voice recognition or embedded cameras, this can lead to serious security and privacy threats.

Issue 8. Home-made devices not implementing industry standards can be produced and these devices can be integrated together with standardized IoT devices.

Issue 9. The dependency of the user to the network service is slowly, but constantly, growing, and this trend has to be expected to continue. In the IoT solutions, this can be especially critical in the case of medical or mission-critical services, where the reliability of the service must be ensured.

More issues can be identified; in this discussion, we tried to identify the most significant potential problems. Table 1 matches the identified issues with their consequences for the system testing processes.

After this initial analysis, let us discuss the IoT quality aspects and techniques, which are currently being researched.

Table 2 Number of papers related to principal categories

Category	Number of papers
Security issues	261
User's privacy and trust issues	43
IoT testbeds	38
Quality assurance and testing techniques	29

3 Related Work

In the current literature, several principal areas dealing with IoT quality can be identified. We can categorize them as the following: (1) security issues, (2) user's privacy and trust issues, (3) reports on IoT testbeds and (4) other quality assurance and testing techniques not related to security, privacy, and particular testbeds. In this section, we summarize these areas.

In our literature survey, we analyzed selected 371 papers related to the categories above from the IEEExplore, ACM Digital Library, and SpringerLink databases. Papers shorter than 4 pages, technical reports, and popular articles were excluded from the analysis. Table 2 summarizes the numbers of papers related to these categories.

In the related literature, **Security issues** are frequently discussed. A number of papers raise the concerns related to security issues, for example [3, 6, 10], and analyze the possible security problems [4, 5]. Moreover, for security testing as a standalone discipline, a number of reports can be found, as an example, we can give [11, 12]. Furthermore, a number of secure architectures on a conceptual and physical level are discussed, for instance [13, 14]. The security area is covered by live publication activity, which reflects on the importance of the issues related to IoT security.

A related topic, user's **privacy and trust** is also being frequently discussed. Concerns are raised [7, 8] and together with that, privacy-aware IoT architectures are being reported [15, 16]. In some of the studies, the privacy and trust topic is overlapping with the security issues, for instance [6, 10].

In the literature, a number of reports on various **IoT testbeds** (or test environments) can be found. Proposed architectures of these testbeds vary from standalone setups [17], distributed architectures [18], or crowd-sourcing based testbeds [19]. Some of the proposals are also based on the simulation of IoT physical devices, e.g. [20], which is a logical step due to the costs of a physical test environment.

The remaining area to discuss is **QA and testing techniques**. This area covers functional testing of IoT solutions, its integration testing, Model-Based Testing and related techniques. Due to the scope of our paper, these reports are the main subject of our interest. Here, we present the more detailed overview.

Several standard-established sub-disciplines of system testing research are spanning to the IoT testing currently. As the initial example, we can give the **Model-Based Testing**. IoT systems are being modeled by a semantic description of IoT services

[21] or by several IoT-specific variants of state machines [22]. Also, UML-based models can be found; for instance, UML class and object diagrams are combined with Object Constraint Language [23]. Alternatively, UML Sequence diagrams with Π-calculus are used [24]. From these models, test cases are generated automatically, which increases the accuracy and coverage of these tests. Closely related to the Model-Based Testing, the **Model Checking** discipline has its representatives in the specific IoT context. To detect possible inconsistencies and defects in IoT models, Computation Tree Logic, CTL [25], δ-Calculus [26] or Temporal Logic of Actions (TLA) formal specification language, based on temporal logic [27] are used. The first representatives of the **run-time verification** of the IoT solutions can be found [28]. In this context, we can also mention representatives of the **IoT reliability models**, combining the hardware and software layer [29, 30] or focusing solely on the software level [31].

As a standalone area, the IoT **protocol testing** can be identified. Variety of the methods is used here, for instance, conformance testing [32], randomness testing [33], statistical verification [34], or formal verification [35]. Previous work related to **IoT usability** testing can be also identified, for instance, an IoT-specific usability testing framework [36]. Several studies can be found discussing the **IoT performance** [37]. Generally, the performance studies focus more on the protocol level, than to the end-to-end performance of the IoT solution from the user's viewpoint.

However, according to the importance of IoT as an emerging technology, more related literature covering the topics of IoT-specific testing and quality assurance can be expected. We discuss this issue later in Sect. 5.

4 The Industry Survey

During the year 2017 we performed structured interviews with ten IoT solution providers, mostly large international companies. The providers varied by the particular IoT business, which included: (1) smart cars, (2) home appliances, (3) smart TVs, (4) and (5) infrastructure for IoT, meaning production of universal IoT devices, from which a final product can be built, (6) R&D, consulting and optimization of IoT solutions and (7)–(10) industrial IoT applications and sensor networks.

Table 3 presents the data related to the question "*Which of the following quality aspects of the IoT solutions do you consider the most challenging?*". The numbers in the header denote the particular IoT provider (the numbers are corresponding to the overview above). The possible answers were 3 to 1, where 3 means the highest possibility. The last column sums the answers; consequently, the discussed issues are sorted from the most significant one.

The issues were specified as follows. **Limited connection** means behavior of the IoT system under limited network connection. **Interoperability** included mutual compatibility of the IoT devices, missing or insufficient standards and a question of proprietary versus internet standards. The **number of configurations** means the number of various configurations and types of the end nodes, making the solution

Table 3 IoT quality issues considered significant

Issue/IoT provider	1	2	3	4	5	6	7	8	9	10	Sum
Limited connection	3	3	2	3	3	3	2	3	3	3	28
Interoperability	3	3	3	3	2	3	2	3	2	3	27
Number of configurations	3	3	3	3	3	2	1	3	3	3	27
Security	3	3	1	2	3	3	2	3	3	3	26
Integration	3	1	2	2	2	3	3	3	3	3	25
Test effort focus	3	1	3	3	2	3	1	2	3	2	23
Performance	3	2	3	2	1	3	2	3	1	2	22
Privacy	3	3	3	2	2	2	1	2	1	2	21
Legislation	2	1	2	3	1	3	1	1	2	1	17

hard to test on all these combinations, in software testing, this effect is called "combinatorial explosion."

Security issues cover various security breach scenarios, where IoT device serve as a weak entry point to the network, possible security breach leading to a personal harm of the user, or security breach leading to a violation of the user's privacy. Here, the area overlaps with the **Privacy**, which also covers possible misuse of collected personal data and reconstruction of user's digital portrait from various data streams. **Integration** issues include challenges how to test interactions of the individual IoT devices and their behavior in the edge cases, this area also relates to the interoperability of the devices.

Test effort focus stood for a challenge, how to determine an efficient and specific test strategy for an IoT solution, which would determine the intensity of testing, test levels, and specific testing techniques. **Performance** issue covered behavior of the IoT solution under possible user traffic peeks and various limited conditions (e.g., a combination of the user traffic peek with a limited network connection). Finally, **Legislation** covered various issues related to the necessity to comply with local legislation, or vague definitions of the implementation rules in this legislation.

Regarding the IoT quality issues considered as significant, the results of the survey presented in Table 3 are relatively balanced; rather than pointing out a clear outlier, the data document, that the mentioned aspects are considered important by the industry representatives. Moreover, IoT quality issues considered significant varied by particular business domain of the IoT solution provider.

As the most significant issues, a behavior of IoT solution on a limited connection, interoperability and problems with a number of various versions and platform variants to test have been pointed out, closely followed by security and integration issues.

5 Discussion

Considering the related literature covering the principal IoT quality areas (Sect. 3, Table 2), a discussion can be made, whether integration, interoperability, platform variants and limited connection problems, shall be covered by the more intense development of IoT-specific testing and quality assurance techniques.

A question can be raised, whether the current software and system testing techniques in their generic form are insufficient to ensure proper testing of the IoT solutions. However, from our feedback from the industry survey (Table 3) as well as from our findings in the initial analysis (Table 1), the conclusion suggests, that this area is rather potential for future research.

In this section, let us further discuss three of these areas: (1) interoperability, (2) behavior of IoT solutions on a limited connection and (3) testing problems caused by a number of various versions and platform variants.

The interoperability of various IoT devices can be addressed by IoT-specific testing methods in two lines. The first line raises the current demands on automation of integration testing and simulation of parts of an IoT infrastructure. Consequences go to the Model-Based Testing discipline. Here, path-based or state-machine-based test case generation techniques can be adapted to the IoT-specific context.

The second line focuses on unit-level integration testing and raises demands to select suitable platform variants, also to generate efficient sets of input testing data for this integration tests. This generates an opportunity for the Constrained Interaction Testing discipline.

Also, the behavior of IoT solution under a limited network connection (or other solution-specific limiting constraints) raises the demands for specific integration and end-to-end testing; also, here, Model-Based Testing discipline could provide more specific methods. A possible approach could be modeling the reliability of the particular network lines in the model of the System Under Test and reflection of these specifics in a generation of special test cases addressing this problem.

Finally, a high number of platform configurations and variants to test is the domain of the Combinational Interaction Testing and Constrained Interaction testing disciplines. IoT-specific models for this problem can be created by modification of the current modeling notations (e.g. Combinational Arrays of Feature Models) to allow generation the test cases efficiently addressing the problem.

Also, overlapping with the interoperability issue, increased demand for integration testing of the IoT solutions and automation of these tests opens an opportunity for further development of integration testing frameworks, to decrease potential maintenance of automated tests, frequently reported as the major drawback of this technology [38]. In the area of front-end based automated testing, the maintenance issues are covered by previous work, e.g. [39–41]. However, this is not the case for the automated integration testing for IoT solutions.

Hence, one of the possible directions here is development of integration testing framework based on unit test framework principles; however, being technically adopted to specifics of the integration test. As an example, we can give higher support for orchestration of an integrated test, more possibilities to chain and execute conditional test steps and more flexible interruption handling of the test flow, all this features also implicitly contributing to decreased maintenance costs of the automated testware.

6 Conclusion

Despite the fact, that IoT represents the major and significant stream in the current technology development, related work addressing the topics of IoT-specific testing methods is rather limited. The industry survey presented in this paper documents the demand of the IoT solution providers for efficient testing and quality assurance methods, developed for IoT specific environment.

During our analysis, we have identified three principal areas, which can be the subject of the further research of IoT-specific testing methods: interoperability testing techniques, techniques for testing of the behavior of the IoT solution under a limited network connection and techniques to efficiently reduce a high number of platform configurations and variants to test. Also, automated integration testing of IoT solutions is one of the prospective streams to be explored further.

IoT-specific Model-Based Testing is one of the suitable candidates to contribute to this area, moreover, Model Checking discipline can explore possibilities of static testing of IoT designs to minimize design errors in IoT solutions. Due to the present intensive research and development work in the IoT technology, we can expect more methods to be developed by the research community; however, currently, these areas represent further research opportunities.

Acknowledgements This research is conducted as a part of the project TACR TH02010296 Quality Assurance System for Internet of Things Technology.

References

1. Kiruthika J, Khaddaj S (2015) Software quality issues and challenges of Internet of Things. In: 2015 14th international symposium on distributed computing and applications for business engineering and science (DCABES). IEEE, pp 176–179
2. Marinissen EJ, Zorian Y, Konijnenburg M, Huang CT, Hsieh PH, Cockburn P, Verbauwhede I 2016. May). Iot: source of test challenges. In: 2016 21th IEEE European Test Symposium (ETS). IEEE, pp 1–10
3. Xu T, Wendt JB, Potkonjak M (2014) Security of IoT systems: design challenges and opportunities. In: Proceedings of the 2014 IEEE/ACM international conference on computer-aided design. IEEE, pp 417–423
4. Bertino E, Choo KKR, Georgakopolous D, Nepal S (2016) Internet of Things (IoT): smart and secure service delivery. ACM Trans Internet Technol (TOIT) 16(4):22
5. Sicari S, Rizzardi A, Grieco LA, Coen-Porisini A (2015) Security, privacy and trust in Internet of Things: the road ahead. Comput Netw 76:146–164

6. Lin H, Bergmann NW (2016) IoT privacy and security challenges for smart home environments. Information 7(3):44
7. Sajid A, Abbas H (2016) Data privacy in cloud-assisted healthcare systems: state of the art and future challenges. J Med Syst 40(6):155
8. Worthy P, Matthews B, Viller S (2016) Trust me: doubts and concerns living with the Internet of Things. In: Proceedings of the 2016 ACM conference on designing interactive systems. ACM, pp 427–434
9. Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M (2015) Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun Surv Tutor 17(4):2347–2376
10. Agrawal V (2015) Security and privacy issues in wireless sensor networks for healthcare. Internet of Things. Springer, User-Centric IoT, pp 223–228
11. Desnitsky V, Kotenko I (2016) Automated design, verification and testing of secure systems with embedded devices based on elicitation of expert knowledge. J Ambient Intell Humaniz Comput 7(5):705–719
12. Fernández-Caramés TM, Fraga-Lamas P, Suárez-Albela M, Castedo L (2016) Reverse engineering and security evaluation of commercial tags for RFID-based IoT applications. Sensors 17(1):28
13. Sicari S, Rizzardi A, Miorandi D, Cappiello C, Coen-Porisini A (2016) A secure and quality-aware prototypical architecture for the Internet of Things. Inf Syst 58:43–55
14. Ashraf QM, Habaebi MH (2015) Autonomic schemes for threat mitigation in Internet of Things. J Netw Comput Appl 49:112–127
15. Wu F, Xu L, Kumari S, Li X (2017) A privacy-preserving and provable user authentication scheme for wireless sensor networks based on internet of things security. J Ambient Intell Humaniz Comput 8(1):101–116
16. Chatzigiannakis I, Vitaletti A, Pyrgelis A (2016) A privacy-preserving smart parking system using an IoT elliptic curve based security platform. Comput Commun 89:165–177
17. Kawazoe H, Ajitomi D, Minami K (2015) A test framework for large-scale message broker system for consumer devices. In: 2015 IEEE 5th international conference on consumer electronics-Berlin (ICCE-Berlin). IEEE, pp 24–28
18. Rosenkranz P, Wählisch M, Baccelli E, Ortmann L (2015) A distributed test system architecture for open-source IoT software. In: Proceedings of the 2015 workshop on IoT challenges in mobile and industrial systems. ACM, pp 43–48
19. Fernandes J, Nati M, Loumis NS, Nikoletseas S, Raptis TP, Krco S, Ziegler S (2015) IoT lab: towards co-design and IoT solution testing using the crowd. In: 2015 international conference on recent advances in Internet of Things (RIoT). IEEE, pp 1–6
20. Giménez P, Molina B, Palau CE, Esteve M (2013) SWE simulation and testing for the IoT. In: 2013 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 356–361
21. Kuemper D, Reetz E, Tönjes R (2013) Test derivation for semantically described IoT services. In: Future network and mobile summit (FutureNetworkSummit), 2013. IEEE, pp 1–10
22. Peischl B (2015) Software quality research: from processes to model-based techniques. In: 2015 IEEE eighth international conference on software testing, verification and validation workshops (ICSTW). IEEE, pp 1–6
23. Ahmad A, Bouquet F, Fourneret E, Le Gall F, Legeard B (2016) Model-based testing as a service for IoT platforms. In: International symposium on leveraging applications of formal methods. Springer, pp 727–742
24. Ren G, Deng P, Yang C, Zhang J, Hua Q (2015) A formal approach for modeling and verification of distributed systems. In: International conference on cloud computing. Springer, pp 317–322
25. Jia Y, Bodanese E, Bigham J (2012) Model checking of the reliability of publish/subscribe structure based system. In: 2012 1st IEEE international conference on communications in China (ICCC). IEEE, pp 155–160
26. Choe Y, Lee S, Lee M (2016) SAVE: an environment for visual specification and verification of IoT. In: 2016 IEEE 20th international enterprise distributed object computing workshop (EDOCW). IEEE, pp 1–8

27. Hillah LM, Maesano AP, De Rosa F, Kordon F, Wuillemin PH, Fontanelli R, Maesano L (2017) Automation and intelligent scheduling of distributed system functional testing. Int J Softw Tools Technol Transfer 19(3):281–308
28. González L, Cubo J, Brogi A, Pimentel E, Ruggia R (2013) Run-time verification of behaviour-aware mashups in the Internet of Things. In: European conference on service-oriented and cloud computing. Springer, pp 318–330
29. Ahmad M (2014) Reliability models for the Internet of Things: a paradigm shift. In: 2014 IEEE international symposium on software reliability engineering workshops (ISSREW). IEEE, pp 52–59
30. Yong-Fei L, Li-Qin T (2014) Comprehensive evaluation method of reliability of Internet of Things. In: 2014 ninth international conference on P2P, parallel, grid, cloud and internet computing (3PGCIC). IEEE, pp 262–266
31. Behera RK, Reddy KHK, Roy DS (2015) Reliability modelling of service oriented Internet of Things. In: 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO) (Trends and Future Directions). IEEE, pp 1–6
32. Xie H, Wei L, Zhou J, Hua X (2013) Research of conformance testing of low-rate wireless sensor networks based on remote test method. In 2013 fifth international conference on computational and information sciences (ICCIS). IEEE, pp 1396–1400
33. Göhring M, Schmitz R (2015) On randomness testing in physical layer key agreement. In: 2015 IEEE 2nd world forum on Internet of Things (WF-IoT). IEEE, pp 733–738
34. Bae H, Sim SH, Choi Y, Liu L (2016) Statistical verification of process conformance based on log equality test. In: 2016 IEEE 2nd international conference on collaboration and internet computing (CIC). IEEE, pp 229–235
35. Silva DS, Resner D, de Souza RL, Martina JE (2016) Formal Verification of a Cross-Layer, Trustful Space-Time Protocol for Wireless Sensor Networks. In: Information systems security. Springer, pp 426–443
36. Wittstock, V., Lorenz, M., Wittstock, E. and Pürzel, F. 2012. A Framework for User Tests in a Virtual Environment. In *Advances in Visual Computing*, 358–367
37. Batalla JM, Gajewski M, Latoszek W, Krawiec P (2015) Implementation and performance testing of ID layer nodes for hierarchized IoT network. In: Asian conference on intelligent information and database systems. Springer, pp 463–472
38. Bures M (2014) Automated testing in the Czech Republic: the current situation and issues. In: Proceedings of the 15th international conference on computer systems and technologies. ACM, pp 294–301
39. Bures M (2015) Framework for assessment of web application automated testability. In: Proceedings of the 2015 conference on research in adaptive and convergent systems. ACM, pp 512–514
40. Bures M (2015) Metrics for automated testability of web applications. In: Proceedings of the 16th international conference on computer systems and technologies. ACM, pp 83–89
41. Bures M (2015) Model for evaluation and cost estimations of the automated testing architecture. In: New contributions in information systems and technologies. Springer, pp 781–787

Using Wi-Fi Enabled Internet of Things Devices for Context-Aware Authentication

Michal Trnka, Filip Rysavy, Tomas Cerny and Nathaniel Stickney

Abstract The increasing spread and adoption of the Internet of Things allows for novel methods to gather information about a user's context, which can be used for enhanced authentication. In this article, we focus on context-aware authentication using information about Wi-Fi networks from a user's wearables or nearables. We propose an additional factor for multi-factor authentication based on the other devices present on the same Wi-Fi network. Devices periodically discover all available peer MAC addresses. During subsequent authentication attempts, the network state is compared to previous network states saved under functionally similar conditions. If the devices on the network change significantly, a flag is raised and further action can be triggered. We also demonstrate the solution as a proof of concept.

Keywords Internet of Things · Security · Wi-Fi · Authentication

1 Introduction

Security is a major concern for all software, and user authentication is a crucial factor in security. There are various methods of enhancing security; one of them is using context [1]. Having the possibility to distinguish a user not only based on their credentials but also using their context as an additional factor brings us new opportunities for improving application security. Context-awareness has not

M. Trnka (✉) · F. Rysavy
Computer Science, FEE, Czech Technical University, Prague, Czech Republic
e-mail: trnkami1@fel.cvut.cz

F. Rysavy
e-mail: rysavfi1@fel.cvut.cz

T. Cerny · N. Stickney
Department of Computer Science, Baylor University, Waco, TX, USA
e-mail: tomascerny@baylor.edu

N. Stickney
e-mail: nathanielstickney@baylor.edu

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_62

only the possibility to enhance security but works in conjunction with personalized applications [2] that improve the user experience as well as allow application owners or operators to target a content precisely for a user. Although the idea of context-aware security is already known [3], few applications use it now because of the various challenges involved.

One of the main challenges is obtaining specific context information that is potentially useful to increase security. The Internet of Things [4] (IoT) provides an opportunity to obtain such useful data. The concept of the IoT is that everyday devices can be interconnected and provisioned to generate, process, and exchange data. This data can be used be used by various applications either to control the devices or simply benefit from them. One beneficial use of this data is improving security. The advantage of using IoT devices is that they can extract data automatically, without any user interaction, so it does not affect the user experience.

In this paper we propose to address the above problem, and demonstrate a method of extracting context from IoT devices as well as using that data for authentication. We focus on devices that users usually carry with them, which have the capability to extract information about context. Those devices can provide us with precise virtual location data, specifically with information about the wireless networks as well as other Wi-Fi and Bluetooth enabled devices either connected to the network or simply in wireless range. The information thus gathered about the user's context can be used as an additional factor during authentication. The idea is to leverage patterns in a user's behavior, as well as the behavior of the environment, measured through the network context. If the device detects roughly the same network context at the same time of day, or the same days of the week, as in previous measurements, we are more likely to authenticate the user. If the user changes their virtual location (network context) significantly, we can ask the user to perform additional authentication steps or apply more restrictive access rules.

This paper describes related work in Sect. 2. In Sect. 3 we describe our proposed method for enhanced authentication using IoT devices. The suggested approach is evaluated in Sect. 4, which describes our experimental implementation and results. In Sect. 5, we summarize our work and develop possible future research directions.

2 Related Work

The idea of context-aware security is more than 15 years old [3, 5, 6]. The first approaches were limited to the context that they could get from communication with the user. Technical limitations generally restricted available contextual information to the history of the interaction, time, approximate location of the user (determined by IP address), and server properties (e.g. CPU load, memory usage). Still, early approaches provided valuable architectures that can be used today with advanced context retrieval methods. The papers cited here describe the possibility to use environmental roles with role-based access control [5], propose a model with four elements—context owner, context provider, context broken and context-aware service [3]—and describe a solution for dynamically calculating properties based on context [6].

The limitations on gathering context information can now be more easily overcome due to the spread of the IoT. Context categorization schemes and their scopes, context acquisition methods, context modelling and representation techniques, and a categorization of context reasoning decision modelling techniques along with how those techniques can be used is described in a comprehensive survey [7].

Numerous context-aware applications using data from the IoT have been described. One paper describes a framework for context-aware sensor search to enumerate and select the best available sensors for a given measurement in the current context [8]. This addresses the common issue of sensor over-proliferation in IoT networks. Another paper from Roggen et al. shows a kit for on-body sensors which recognize context and user activity [9]. The kit allows easy access to contextual information, with the ability to add hardware and modify the kit's software.

Significant effort has been put into using data acquired through Wi-Fi networks to enhance security. Various methods exists to acquire user's precise location based on Wi-Fi signals [10, 11]. They do not require user cooperation, or even that the user wear or carry a device. No special equipment is needed beyond a standard Wi-Fi router to measure the user's position. The user's location is determined based on the properties of the signal (e.g. amplitude, strength, etc.). The median precision is less than 30 cm. Cong et al. extends the idea by leveraging the Wi-Fi signal from various IoT devices [12].

The research topics most relevant to our proposal are those which enhance security using contextual information extracted by IoT devices. Agadakos et al. presents a service which determines whether a user can be in a given location based on their previous locations and time elapsed, among other factors [13]. The service is then used as additional factor during authentication. Shahzad et al. describes a solution using biometric information to continuously authenticate users [14]. Shone et al. proposes authentication using the user's digitalized memories [15]. Users would authenticate themselves by remembering previously digitalized events and giving details such as dates and times, places, people or pets, devices, habits, and audio or ownership recognition. Said et al. applies a slightly different approach. The authors design a context-aware trust management system [16] that calculates trust among devices in the network dynamically based on the context of the devices.

3 Proposed Method

To tackle the problem of gathering extra contextual information, we propose to use communicating IoT devices as our data source. If the device is a sensor, or contains at least one sensor, it can provide us with valuable environmental context. Also, by definition every IoT device is connected to the Internet and very often this is done through a computer network. In this case it can provide us with network information.

The context acquired through the device needs to be tied to a specific user, but the relationship between users and devices can exist in various forms in terms of mapping devices to users and vice versa. The situation most clearly applicable for our purposes is when a device is owned permanently by a single user, and that user wears, carries, or otherwise has the device close to them (e.g. is personally using it) the majority of the time. Given those constraints, we can easily tell that the context acquired by the device belongs to the user in question. In our proposal we will consider only this constrained set of devices and not the devices that can belong to, or provide data for, multiple users.

In the existing literature, the most common use of IoT sensors is to determine a user's physical, environmental context [7]. We will instead focus on the user's virtual location and their context regarding this virtual position. Information about the network the device is connected to (or the connections that are available) provides a contextual picture of the virtual location. The initial data points for available networks are the strength of the signals, MAC addresses of the routers (BSSID), and names of the networks. If the device is actually connected to a network, we can further retrieve subnet mask, gateway address, DHCP and DNS servers address if available, and the IP and MAC addresses of others devices connected to the network.

To fully leverage the information gathered about the network requires a significant amount of contextual data gathered over extended periods of time in multiple distinct physical locations and across multiple different networks. Such a large data set might be best analyzed using data science techniques to determine the correct "security risk" level for a given context change. Unfortunately, we do not have such a data set available yet. Therefore, in this paper we describe an initial, intuitive method to determine whether a user's network context requires additional actions (e.g. additional factor for authentication or authorization) or not.

Our idea is based on determining "stable" devices. Stable devices are those that are typical and not changing for a given network in a given time. Also, they must not be prevalent in the other networks. The reasoning behind this constraint is that every network should posses various distinct stable devices—router, printer, etc. Those devices should be stable for any time of a day and day of a week. Personal devices present on the network should also present a roughly similar footprint during identical times of the day, as people tend to follow regular schedules.

During every action the user performs, we store the network context. Because the network scan is time consuming, we gather the network data in regular time intervals, preferably shorter than an hour, or when a network changes. The information is stored for further use and comparison. Based on the historical values, we determine stable devices. Those are the devices that appear at the same times during the day on the network.

We also need to determine devices that are tied with the user—his secondary devices, or devices owned by his family members and immediate coworkers. Because these devices may appear in multiple virtual locations along with the user's device, they are not suitable for determining the virtual location itself. They are excluded from our determination of stable devices because they are not network specific and we expect that their presence is connected more to a user than to contextual network events.

During a user's interaction with the application, we can compare the information about the stable devices with the immediate state of the devices on the network. This might be combined with other contextual information, like exact GPS location. If the devices present at the given time differ greatly from the devices present at the same time on previous days, we flag it as suspicious. In order to do so we compare MAC addresses with the record for roughly similar times in the history. The three latest states are used as a benchmark for current state. Based on preliminary experiments, we set a threshold that 70% of devices present in the historical network context should correspond to the current network context, or the current context is marked as suspicious and further steps can be taken—additional authentication factor can be invoked, some access rights can be revoked, etc.

Theoretical limitations of the method described above include the need to determine the network information. This information can't be retrieved using a normal web application. It would limit the usage to native applications (desktop or mobile—Android or iOS). This means a different front-end application for every supported operating system. Also, there is a question about the user's data privacy and whether users would like to share information about their network context and how the information would be stored and used, to prevent abuse or accidental disclosure.

4 Evaluation

Because of the lack of a common development platform for IoT devices, we have chosen to implement our proof of concept for the Android operating system. We created an application that detects all other devices on the same network. That application is used for testing the feasibility of our proposal. Source code and executable of the application can be found in the following repository: https://bitbucket.org/frysavy/context-acquisition-app.

Initially, the IP address of the device and subnet mask of the network is determined. With this information, we have the range of IP addresses in the subnet. Then, we ping every address in the range with a timeout set for 1000 ms. For those that reply, we match IPs to their MAC address from the ARP table at/`proc/net/arp`. The results are stored in the JSON file for further processing.

Without parallel processing, the network scan takes a noticeable amount of time on large networks. Another drawback of this method is that devices on the network must respond to the ICMP echo requests. Some firewalls block such communication, and some devices ignore it as well. Also, the devices need to be on same physical network or on networks between which the routers allow propagation of ICMP echo requests. For this reason, devices on Wi-Fi networks on the same subnet, but separated between the 2.4 and 5 GHz channels, might not know about each other.

We performed multiple measurements of the same network at identical times on the same physical spot, on various days. As a testing network, we chose Baylor

Table 1 Benchmark creation

	Total devices found	Overlap from previous days
Day 1	447	
Day 2	434	77
Day 3	469	46
Day 4	484	38

Table 2 Evaluation results—measurements at various places compared to benchmark and the result of the authentication check

	Total amount of device	Number of stable devices	Percentage of stable devices (%)	Result
Benchmark		38		
Weekday 1	450	31	82	True
Weekday 1	463	30	79	True
Weekday 1	461	31	82	True
Weekday 1	441	28	74	True
Weekday 1	432	31	82	True
Weekend	20	15	32	False
Apartment	3	0	0	False

University's Wi-Fi network as it provided great variability and large numbers of users and devices. We performed three measurements during regular weekdays, always starting at 10 AM. The goal of this was to determine stable devices on the network. During our measurements there were over 400 devices active every day. With two days, there were 77 common MAC addresses. When we added the third day, it decreased to 46 MAC addresses and on the fourth day it decreased to final 38 MAC addresses that we consider stable. The data used for this benchmark are illustrated in Table 1.

With the benchmark set, we took seven different measurements to verify our method. Five of them occurred on a weekday at 10 AM., at the same physical location as the previous measurements used for the benchmark. Between 74 and 82% of the benchmark devices were found during those measurements, which is inside our threshold. Two control measurements were performed to verify that our method can detect a change. One was performed at 10 AM. during weekend at the same spot. The second one was taken at 9 AM. at one author's apartment. During both of them, the authentication factor returned false. The results are represented in Table 2. It shows the numbers of devices on the network for every network evaluation, number of MAC addresses that match those in benchmark set, and the match percentage.

Evaluation across a large number of situations with more users is needed to definitely confirm our method not only as correct but also as practically usable. However, the preliminary results show that the method provides valid results and can enhance authentication.

5 Conclusion

The increasing spread of IoT devices brings new possibilities to obtain a context. Contextual information can be used for various improvements for applications including security. In this paper we focused on context-aware authentication. We have described a method using IoT devices to gather network information and the process of using this information as an additional factor for multi-factor authentication. The solution provides additional security while preserving the user's convenience by automatically observing the state of the network and comparing it to the states of the network in the past.

We demonstrated the feasibility of our approach on a small but essential proof of concept using a single IoT device. The results are positive and indicate that the approach we have chosen can provide valid results and increase the security of applications. Further evaluation is needed to definitely confirm the strengths of the work.

In future work we want to focus on creation of a more extensive, intensive, and detailed analysis of the method. Further, we want to find a method of dynamically linking a user to a IoT devices that are close to him, so we can use them to extract context. Finally, we want to experiment with using the authentication method not only for users, but also for single IoT devices.

Acknowledgements This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/234/OHK3/3T/13.

References

1. Abowd GD, Dey AK, Brown PJ, Davies N, Smith M, Steggles P (1999) Towards a better understanding of context and context-awareness. In: Gellersen HW (ed) Handheld and ubiquitous computing. Heidelberg, Springer, Berlin Heidelberg, Berlin, pp 304–307
2. Hong J, Suh EH, Kim J, Kim S (2009) Context-aware system for proactive personalized service based on context history. Expert Syst Appl 36(4):7448–7457
3. Hulsebosch RJ, Salden AH, Bargh MS, Ebben PWG, Reitsma J (2005) Context sensitive access control. In: Proceedings of the tenth ACM symposium on access control models and technologies. SACMAT'05, New York, NY, USA. ACM, pp 111–119
4. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15):2787–2805
5. Covington MJ, Long W, Srinivasan S, Dev AK, Ahamad M, Abowd GD (2001) Securing context-aware applications using environment roles. In: Proceedings of the sixth ACM sympo-

sium on access control models and technologies. SACMAT'01, New York, NY, USA. ACM, pp 10–20

6. Bhatti R, Bertino E, Ghafoor A (2004) A trust-based context-aware access control model for web-services. In: Proceedings of the IEEE international conference on web services. ICWS'04, Washington, DC, USA. IEEE Computer Society, 184
7. Perera C, Zaslavsky A, Christen P, Georgakopoulos D (2014) Context aware comput-ing for the internet of things: a survey. IEEE Commun Surv Tutor 16(1):414–454
8. Perera C, Zaslavsky A, Christen P, Compton M, Georgakopoulos D (2013) Context-aware sensor search, selection and ranking model for internet of things middleware. In: 2013 IEEE 14th international conference on mobile data management, vol 1, June 2013, pp 314–322
9. Roggen D, Bchlin M, Schumm J, Holleczek T, Lombriser C, Trster G, Widmer L, Majoe D, Gutknecht J (2010) An educational and research kit for activity and context recognition from on-body sensors. In: 2010 international conference on body sensor networks, June 2010, pp 277–282
10. Wang Y, Liu J, Chen Y, Gruteser M, Yang J, Liu H (2014) E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In: Proceedings of the 20th annual international conference on mobile computing and networking. MobiCom'14, New York, NY, USA. ACM, pp 617–628
11. Adib F, Kabelac Z, Katabi D, Miller RC (2014) 3d tracking via body radio reflections. In: Proceedings of the 11th USENIX conference on networked systems design and implementation. NSDI'14, Berkeley, CA, USA. USENIX Association, pp 317–329
12. Shi C, Liu J, Liu H, Chen Y (2017) Smart user authentication through actuation of daily activities leveraging wifi-enabled IoT. In: Proceedings of the 18th ACM international symposium on mobile ad hoc networking and computing. Mobihoc'17, New York, NY, USA. ACM, pp 5:1–5:10
13. Agadakos I, Hallgren P, Damopoulos D, Sabelfeld A, Portokalidis G (2016) Location-enhanced authentication using the IoT: because you cannot be in two places at once. In: Proceedings of the 32nd annual conference on computer security applications. ACSAC'16, New York, NY, USA. ACM, pp 251–264
14. Shahzad M, Singh MP (2017) Continuous authentication and authorization for the internet of things. IEEE Internet Comput 21(2):86–90
15. Shone N, Dobbins C, Hurst W, Shi Q (2015) Digital memories based mobile user authentication for IoT. In: 2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing, Oct 2015, pp 1796–1802
16. Saied YB, Olivereau A, Zeghlache D, Laurent M (2013) Trust management system design for the internet of things: a context-aware and multi-service approach. Comput Secur 39:351–365

A Hybrid Deep Q-Network for the SVM Lagrangian

Chayoung Kim and Hye-young Kim

Abstract The setting hyperparameters in the support vector machine (SVM) is very important with regard to its accuracy and efficiency. In this paper, we employ a novel definition of the reinforcement learning state, actions and reward function that allows a deep Q-network (DQN) to learn to control the optimization hyperparameters for the SVM deep neural networks by supervised Big-Data. In this framework, the DQN algorithm with experience replay is based on the off-policy reinforcement learning for the expected discounted return of rewards, or q-values, connected to the actions of adjusting the hyperprameters in the SVM. We propose the two deep neural networks, one with the SVM and the other with Q-network (DQN). The SVM deep neural networks learns a policy for the optimization hyperparameters, but differ in the number of allowed actions. The SVM deep neural networks trains the hyperparameters of the SVM simultaneously such as the Lagrangian multiplier. The proposed algorithm is called a Hybrid DQN combined with SVM deep neural networks. This algorithm could be considered as the classifier in the real-world domains such as network anomalies in the distributed server loads, because the SVM is suitable for the application in a classification, especially for the one-against-the others. Algorithm comparisons show that our proposed algorithm leads to good optimization of the Lagrangian multiplier and can prevent overfitting to a certain extent automatically without human system designers. In terms of the classification performance of the proposed algorithm can be compared to the original LIBSVM with no controls of the hyperparameters.

Keywords SVM deep neural networks · Network anomalies in distributed server loads · Hybrid deep Q-Network reinforcement learning · Hyperprameters

C. Kim
Kyonggi University, 154-42 Gwanggyosan-Ro, Yeongtong-gu, Suwon,
Gyeonggi, South Korea
e-mail: kimcha0@kgu.ac.kr

H. Kim (✉)
Hongik University, 2639 Sejong-Ro, Jochiwon-Eup, Sejong, South Korea
e-mail: hykim@hongik.ac.kr
URL: http://nglab.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_63

1 Introduction

One of the supervised machine learning (ML), the support vector machine (SVM) has been widely used in many applications, such as the decision-making application, forecasting malaria transmission, liver fibrosis diagnosis, and pattern classification [1]. The SVM has the significant tradeoff between the minimum training set error and the maximization of the margin based on the Vapnik-Chervonenk' theory and the structural risk minimization principle [1]. The SVM is a convex quadratic programming algorithm with which it is possible to find the global rather than the local optima. However, when it comes to SVM deep neural networks, the setting of the hyperparameters for the SVM deep neural networks plays a more significant role than the SVM without neural networks, which includes the penalty parameter and the smoothness parameter of the radial-based function. The penalty parameter maintains the balance between the fitting error minimization and model complexity. The smoothness parameter of the kernel function is used to determine the nonlinear mapping for the high-dimensional feature space [2]. In general, the redundant features of the SVM without neural networks usually makes the designed overfitting the training data. In general, an effective feature selection can tackle the cure-of dimension problem as well as decrease the computation time. So, there have been some researches for the hyperparameters of the SVM including the grid search [2] under exponentially growing sequences for the penalty and the smoothness parameters. However, Deep reinforcement learning (RL) algorithms [3–7] have been applied in a range of challenging domains including the controlling of hyperprameters without the penalty and the smoothness parameters, especially for the deep neural networks. Like previous research [2], we focus on the hyperparamters, Lagrangian multiplier α_i without the penalty and the smoothness parameters because Lagrangian multiplier is significant for the loss function of the SVM deep neural networks.

Therefore, we employ a deep Q-network (DQN) [8–11] to learn to control the optimization hyperparameters for the SVM deep neural networks based on the RL. The off-policy DQN algorithm with experience replay is based on the maximum entropy reinforcement learning for the expected discounted return of rewards, or q-values, connected to the actions of adjusting the hyperprameters in the SVM deep neural networks. Our two deep neural networks, which is one with the SVM and the other with Q-network (DQN). The DQN learns a policy similar for the optimization hyperparameters and the SVM deep neural networks algorithm to train hyperparameters of the Lagrangian multiplier. This algorithm could be considered as the classifier in the real-world domains such as network anomalies in distributed server loads, because the SVM is suitable for the application in a classification, especially for the one-against-the others SVM. Our goal is to maximize the Lagrangian multiplier through gradient-based algorithms, which are effective for neural network optimization. At each iteration, we extract information about the objective to form a support feature vector. The support feature vector is the input to a DQN and the output is the expected discounted return of rewards, or q-value, connected to the action of increasing, decreasing, or preserving the Lagrangian multiplier. By using the DQN in combination with the SVM deep neural networks with a gradient-based optimization routine to iteratively adjust the hyperparameter, which is the concept for Hybrid DQN combined with SVM deep neural networks. Most of RL studies are based on the weak supervision in the form of a reward given. However, we consider those of supervised ML with SVM deep neural networks instead. And like previous research [11], we uses an off-policy DQN, which combines exploitation and exploration. This proposed algorithm can exploit the supervised deterministic optimization in the first, the SVM deep neural networks and better solutions in the second, the DQN for the first one. By the algorithm comparisons, we can show that our algorithm leads to good optimization of the hyperparameters and can prevent overfitting to a certain extent automatically without human system designers in the SVM deep neural networks.

2 Materials

2.1 Data

KDDCup 1999 [12] dataset has been widely used for anomaly detection methods. Its training dataset is composed of 4,900,000 single connection vectors that contain 41 features and they are labeled as either normal or an attack. Attack labels (types) fall in one of the following four categories:

- *Denial of Service Attack (DoS)*: This is an attack that intends for making computing or memory resources too busy.
- *User to Root Attach (U2R)*: This attack type starts with access to a normal user account on the system, and then tries to exploit vulnerability to get local access.
- *Remote to Local Attack (R2L)*: This attack occurs when an attacker, who does not have an account on the system, sends packets to the system over a network.
- *Probing*: This is an attempt to gather information about computer networks for the purpose of circumventing its security controls.

3 The Proposed Algorithm

In this section, we propose a hybrid deep Q-network algorithm that can automate one of SVM hyperparamters, Lagrangian Multiplier for stochastic gradient descent (SGD) [7] based machine learning (ML) algorithms. Our approach is especially for SVM deep neural networks. Most of ML algorithms need to train a model with some parameters ω by minimizing a loss function f defined over a set X of training examples:

$$\omega* = \mathrm{argmin}_{\omega} \cdot f_{\omega}(\mathrm{X}).$$

At every time step, updating parameters by gradients is the standard approach for the loss function minimization called as gradient descent (GD).

$$\omega^{t+1} = \omega^{t} - \lambda^{t}\nabla f^{t},$$

where λ^{t} is the learning rate at the step time t, and $\nabla\ f^{t}$ is the local gradient of f at ω^{t}.

At a step, we can use the whole batch of the training data, a mini batch of tens/hundreds or a random sample. We can observe that the performance of SGD based methods is quite sensitive to the choice of the hyperparamters of SVM deep neural networks for non-convex loss function f. Generally, f is usually non-convex with respect to the parameters ω in many ML algorithms, especially for deep neural networks. In terms of SVM deep neural networks, we can fine the optimal separating margin by the gradient descent method. Based on the inequality constrained Wolfe Dual form, we can rewrite the SVM as follows,

$$\text{Maximize}\{\sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i=1}^{l} \alpha_i\alpha_j y_i y_j x_i x_j\},$$

$$C \geqq \alpha_i \geqq 0, \quad i = 1, \ldots, l, \qquad \sum_{i=1}^{l} \alpha_i y_i = 0,$$

where C is a penalty value, α is the Lagrangian multiplier.

We aim to learn the Lagrangian multiplier, one of hyperparameters using the deep Q-network (DQN) that can automatically control at each step. Figure 1 shows our automatic Lagrangian multiplier, which adopts the DQN framework. The basic idea is that at each step, given the current model ω^{t} and training sample x, the DQN is used to take an action (the Lagrangian multiplier α_{i}). The proposed algorithm is model-free. It solves the DQN task directly using samples with probability ϵ. That is off-policy. It learns about the greedy strategy α_{max}, while following the adequate exploration of the state space [11]. The ϵ-*greedy* [10] strategy is for selection of the good actions. We describe the details of our algorithm in the following subsections.

3.1 The Deep Q-Network (DQN)

The DQN in RL determines the Lagrangian multiplier for the SVM deep neural networks machine learning. There is the proposed algorithm in Fig. 1 based on the current model, training data, and historical information during the training process. Note that ω^{t} could be of huge dimensions. If DQN takes all of those parameters, its computational complexity would dominate the complexity of the primary algorithm. Therefore, we propose to use a function $\chi(\cdot)$ to process and yield a compact vector s^{t} based on SVM deep neural networks as the input of the DQN. In this proposed architecture, the DQN learns the Lagrangian multiplier from the off-policy and shows

```
Algorithm: A hybrid DQN for Lagrangian multiplier of SVM deep neural networks
Require: Training step T, training set X, loss function $f$, state function $\chi$, discount factor $\gamma$, Replay memory $D$
Ensure: Model parameters $\omega$, value parameters $\rho$ of the deep Q-network, and the action, Lagrangian multiplier $\alpha$
----------------------------------------------------------------
Initial parameters $\omega_0, \rho_0,$ $D$ to capacity $N$
for $t$=0, …, $T$ do
    Sample $x_i \in X$, $i \in 1, \ldots, N$ .
    Extract state vector: $s_i^t = \chi(\omega^t, x_i)$.
    Computes Lagrangian multiplier $\alpha_i^t = \pi_\theta(s_i^t), LM(\alpha_i^t)$
    Otherwise select a random action $\alpha_i^t$ with probability $\epsilon$
    If $LM(\alpha_i^t) > LM(\alpha_{max})$,
        Save $LM(\alpha_i^t)$ as a max,
        Computes the optimized weight vectors, $w=\sum_{i=1}^{N} \alpha_i\, y_i\, x_i$ ,
        Computes a bias, $b = 1\text{-}w^T x_i, y_i=1$
    Compute $\bigtriangledown f^t(x_i)$.
    Update $\omega$: $\omega^{t+1} = \omega^t - \alpha^t_i \bigtriangledown f^t(x_i)$
    Set $r^t = f^t(x_i) - f^{t+1}(x_i)$                    // Update Q-network
    Extract state vector: $s_i^{t+1} = \chi(\omega^{t+1}, x_i)$.
    Compute $Q\rho\,(s_i^{t+1}, \pi_\theta(s_i^{t+1}))$, $Q\rho(, \alpha_i^t)$
    Store transition $(s_i^t, \alpha_i^t, r^t, s_i^{t+1})$ in $D$
    Sample random minibatch of transitions $(s_i^t, \alpha_i^t, r^t, s_i^{t+1})$ from $D$
    Set $\delta^t = r^t + \gamma Q\rho\,(s_i^{t+1}, \pi_\theta(s_i^{t+1})) - Q\rho(s_i^t, \alpha_i^t)$
          $= r^t$                                       // With a random
    Update $\rho$ by $\nabla\rho = \delta^t \nabla\rho Q\rho(s_i^t, \alpha_i^t)$
end for
return
```

Fig. 1 The proposed a hybrid deep q-network for SVM Lagrangian

the automatic Lagrangian multiplier. Following the practice in RL, we call $\chi(\cdot)$ the state function, which takes ω^t and the training data x as inputs:

$$s^t = \chi(\omega^t, X) :$$

Then the DQN $\pi_\theta(\cdot)$ parameterized by θ yields an action α^t, such as the Lagrangian multiplier:

$$\pi_\theta\left(s^t\right) = \alpha^t$$

where the action $\alpha^t \in R$ is a continuous value. The DQN has to learn the input parameters to output a good action. The goal of the proposed algorithm is to find a good action for the Lagrangian multiplier, α^t to ensure that a good model can be learnt eventually by the SVM deep neural networks. For this purpose, the DQN needs to output a good action α^t at state s^t so that finally a low training loss $f(\cdot)$ can be achieved. In DQN, $Q_\pi(s, \alpha)$ is often used to denote the long term reward of the state-action pair (s, α) while following the policy π for Lagrangian multiplier. In our proposed algorithm, $\mathrm{Q}_{\pi}(s^{t}, \alpha^{t})$ indicates the accumulative decrement of training loss starting from step t. We define the immediate reward at step t as the one step loss decrement:

$$r^{\mathrm{t}} = f^{\mathrm{t}} - f^{\mathrm{t+1}}.$$

The accumulative value $\mathrm{R}_{\pi}^{\mathrm{t}}$ of policy π at step t is the total discounted reward from step t:

$$\mathrm{R}_{\pi}^{\mathrm{t}} = \sum\nolimits_{k=t}^{T} \gamma^{\mathrm{k-t}} r\left(s^{t}, \alpha^{t}\right)$$

where $\gamma \in (0; 1]$ is the discount factor. Given that the states and actions are uncountable in our proposed algorithm, the DQN uses a parametric function $\mathrm{Q}\rho\ (s, \alpha)$ with parameters ρ to approximate the value function of DQN, $\mathrm{Q}_{\pi}(s, \alpha)$. The action at each time step is chosen based on the principle of exploration versus exploitation. Exploitation takes advantage of the information already garnered by the DQN while exploration encourages random actions to be taken in prospect of finding a better policy [10, 11]. We employ a *ϵ-greedy* policy which chooses the optimal action w.r.t the DQN's q-values with probability $1 - \epsilon$ and randomly otherwise. Q-learning is an off-policy procedure because it follows a non-optimal policy (with probability ϵ a random action is taken) yet makes updates to the optimal policy.

3.2 The Training the Hybrid DQN

The DQN with experience replay method has the parameters ρ, which is updated at each step using TD learning. The DQN learns how to minimize the loss function through repeated attempts and a trade-off between finding a good Lagrangian multiplier and exploring the space. Restricting the number of time steps reflects real world applications where there are computational and time constraints and exploit a priori knowledge of the objective function based on the SVM deep neural networks. With an experience consisting of a $(s_{i}^{t}, \alpha_{i}^{t}, r^{\mathrm{t}}, s_{i}^{t+1})$ tuple for some time step in replay memory D of experience ϵ like [11]. Instead of updating the DQN with only the most recent experience, a subset $\mathrm{S} \subset \epsilon$ of experiences are drawn from memory and used as a mini-batch to update the DQN. More precisely, the critic is trained by minimizing the square error between the estimation $\mathrm{Q}\rho(s^{t}, \alpha^{t})$ and the target y^{t}:

$$y^{t} = r^{t} + \gamma \mathrm{Q}\rho(s^{t+1}, \alpha^{t+1}) :$$

The TD error is defined as:

$$\begin{aligned} \delta^{t} &= y^{t} - \mathrm{Q}\rho(s^{t}, \alpha^{t}) \\ &= r^{t} + \gamma \mathrm{Q}\rho(s^{t+1}, \pi_{\theta}\left(s^{t+1}\right)) - \mathrm{Q}\rho(s^{t}, \alpha^{t}) \end{aligned}$$

The weight update rule follows the off-policy deterministic DQN. That means that the proposed algorithm learns the Lagrangian multiplier in the way of exploitation and exploration. The gradients of critic network are:

$$\nabla \rho = \delta^t \nabla \rho Q \rho(s^t, \alpha^t)$$

The proposed algorithm can output the action with the largest Q value of deep Q network at state s^t, i.e., $\alpha* = \arg\max_{\alpha} Q\rho(s^t, \alpha)$. In terms of SVM deep neural networks, the maxim Lagrangian multiplier is that $\alpha* = \arg\max_{\mathrm{LM}(\alpha)} Q\rho(s^t, \alpha)$. Mathematically,

$$\nabla \theta = \nabla_{\theta\pi\theta}(s^{t+1}) \nabla_{\alpha} Q\rho(s^{t+1}, \alpha^{t+1}) | \alpha = \pi_{\theta}(\mathrm{s}).$$

3.3 The Algorithm Comparison

In this section, we would like to compare the proposed hybrid DQN with SVM deep neural networks and the SVM without automatic Largrangian multiplier. In Fig. 1, there is our algorithm which is shown that we sample an example, extract the current state vector, and compute Lagrangian multiplier using DQN and sample another example to update the DQN. In this whole processes are based on off-polity. First, in the proposed algorithm, we utilize a technique known as experience replay [10, 11] where we store the agent's experiences at each time-step (s_i^t, α_i^t, r^{t}, s_i^{t+1}) tuple in replay memory *D*. We apply Q-learning updates, or mini batch updates, to samples of experience in D, drawn at random from the pool of stored samples like [10, 11]. After performing experience replay, the agent selects and executes an action according to a ϵ*-greedy* policy. Learning directly from consecutive samples is inefficient, due to the strong correlations between the samples. That means randomizing the samples breaks these correlations and therefore reduces the variance of the updates like [10, 11]. And second, we use one example (e.g., $\mathrm{x_i}$) for the model and the DQN update, but a different example (e.g., $\mathrm{x_j}$) for the SVM deep neural networks. Doing so we can avoid that the SVM without automatic hyperparameters will overfit on some hard examples and can improve the generalization performance of the proposed algorithm on the test set. We can exploit the a-prior knowledge-based advantages by the input data, KDDCup 1999 [12] dataset, which is widely used for anomaly detection with 4 categories based on supervised mode. When it comes to machine learning, the network weight are convex. However, in terms of the neural networks, the weight are non-convex. Based on the SVM without neural networks, it can be solvable by quadratic programming. That means we cannot exploit easy example in a classification task. Therefore, we should consider more difficult to be classified correctly with SVM deep neural networks. In the case of hard examples, the gradient descent will be large and the hyperparameters are not tractable by the human system designs. In other words, this hard example will greatly change the model, while it is not a good representative of its category and the hypeprameters should not pay much attention to it. If we input the same example to both the DQN and the SVM deep neural networks, they will make the model to change a lot to fit the example, while resulting in oscillation of the training. With different example inputs into the both, it is very likely that the DQN will find that the gradient descent direction of the example with the inconsistent training example and thus optimize the Lagrangian multiplier. More precisely, the update of ω is based on x_i and the Lagrangian multiplier by the DQN, while the training target of the DQN is to maximize the output on x_j.

4 Conclusion

We have suggested the algorithm a hybrid deep Q-network (DQN) with replay memory for SVM Lagrangian. The DQN based on the RL to learn to control the optimization hyperparameters for the SVM deep neural networks. The off-policy DQN with the experience replay is based on the expected discounted return of rewards, or q-values, connected to the actions of adjusting the hyperprameters in the SVM deep neural networks. Our two deep neural networks, which is one with the SVM and the other with Q-network (DQN). The proposed algorithm could be considered as the classifier in the real-world domains such as network anomalies in distributed server loads, because the SVM is suitable for the application in a classification, especially for the one-against-the others SVM. In our proposed algorithm, we consider those of supervised machine learning (ML) with SVM deep neural networks instead of week versions given by a reward of reinforcement learning only. With the DQN by learning hyperparameters for ML algorithms, we propose to feed different training examples to the DQN and the SVM deep neural networks, which improve the generalization performance of the learnt supervised ML algorithm. By algorithm comparisons, we can find out more difficult to be classified correctly with the SVM deep neural networks. In the case of hard examples, the gradient descent will be large and the hyperparameters given by the DQN are not tractable by the human system designs. If we input the different example to both the DQN and the SVM deep neural networks, they will make the model not to be over fitted with inconsistent training examples. That means we can show that our algorithm leads to good optimization of Lagrangian multiplier of hyperparameters and can prevent overfitting to a certain extent automatically without human system designers in the supervised ML algorithms.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. 2016RIA2B4012386).

References

1. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machine. Mach Learn 46:389–422
2. Chao C-F, Horng M-H (2015) The construction of support vector machine classifier using the firefly algorithm. Comput Intell Neurosci Arch 2015(2)
3. Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3(1):9–44
4. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, vol 1. MIT press Cambridge
5. Sutton RS, McAllester DA, Singh SP, Mansour Y (1999) Policy gradient methods for reinforcement learning with function approximation. In: NIPS, vol 99, pp 1057–1063 (1999)
6. Sutton RS (1984) Temporal credit assignment in reinforcement learning. Doctoral Dissertation
7. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw Mach Learn 4(2)
8. Silver D, Lever G, Heess N (2014) Deterministic policy gradient algorithms. In: ICML'14 proceedings of the 31st international conference on international conference on machine learning vol 32, pp 387–395
9. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Driessche GVD, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489
10. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. In: NIPS
11. Hansen S (2016) Using deep q-learning to control optimization hyperparameters. ArXiv
12. Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the KDD CUP 99 dataset. In: Proceedings of the 2009 IEEE symposium on computational intelligence in security and defense applications (CISDA 2009), pp 53–58

An Actor-Critic Algorithm for SVM Hyperparameters

Chayoung Kim, Jung-min Park and Hye-young Kim

Abstract There have been a great deal of researches based on deep neural networks for the network anomalies in the areas of network server workloads. We focus on deep neural networks to deal with huge server loads of network anomalies in a distributed MMOGs (massively multiplayer online games). We present an approach to training deep neural networks to control a loss function optimization, such as hyperparameters, using actor-critic methods from reinforcement learning (RL). Deep RL algorithms have been demonstrated on a range of major challenges: brittle convergence properties such as hyperparameter tuning. We propose an algorithm to automatically optimize the one of hyperparameters, Lagrangian multiplier of the support vector machines (SVM) deep neural networks using actor-critic algorithms. The setting of hyperparameters in the SVM deep neural networks is very important with regard to its accuracy and efficiency. Therefore, we employ the actor-critic algorithm to train Lagrangian multiplier of the SVM deep neural networks. We train a policy network called actor to decide the Lagrangian multiplier at each step during training, and a value network called critic to give feedback about quality of the decision (e.g., the goodness of the Lagrangian multiplier given by the actor) that the actor made. Algorithm comparisons show that our algorithm leads to good optimization of Lagrangian multiplier and can prevent overfitting to a certain extent automatically without human system designers.

C. Kim
Kyonggi University, 154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon,
Gyeonggi, Korea
e-mail: kimcha0@kgu.ac.kr

J. Park · H. Kim (✉)
Hongik University, 2639 Sejong-ro, Jochiwon-eup, Sejong, Korea
e-mail: hykim@hongik.ac.kr
URL: http://nglab.kr

J. Park
e-mail: bjmbam12@naver.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_64

Keywords SVM deep neural networks · Massively multiplayer online games (MMOGs) · Network anomalies · Actor-Critic reinforcement learning Hyperparameters

1 Introduction

Deep reinforcement learning (RL) algorithms have been applied in a range of challenging domains, from games [1–3]. The combination of RL and deep neural networks holds the promise of automating a wide range of decision making. However, in terms of real-world domains, such as network anomalies in the distributed architecture of massively multiplayer online games (MMOGs) in Big-Data, there has been hampered by some major challenges. One of those challenges is brittleness with respect to their hyperparameters such as loss functions and other settings, which should be set carefully for achieving good results. These challenges severely limit the applicability of deep RL to real-world tasks. Although there are some empirical suggestions to guide how to adjust the hyperparameters over time in training, it is still a difficult task to find a good policy to adjust the hyperparameters, given that good policies are problem specific and depend on implementation details of a supervised machine learning (ML) algorithm [4, 5]. With regard to supervised ML, we focus on support vector machine (SVM) because the SVM have been widely used in many applications, including the decision-making application, forecasting malaria transmission, liver fibrosis diagnosis, and pattern classification [6]. In general, the redundant features of the classifier usually significantly adjust the penalty parameters and the smoothness parameter, which plays significant roles for the Lagrangian multiplier of the loss function [6]. This means that we need to try many times and adjust the hyperparameters manually to accumulate knowledge about the problem. However, human design often needs domain knowledge about the target problems, which is inefficient and difficult to scale up to the real-world domains. Therefore, we decided to adjust automatically the Lagrangian multiplier without the penalty parameters and the smoothness parameter. This is exactly the focus of this work and we aim to automatically learn the Lagrangian multiplier based on SVM deep neural networks without human design. In terms of control of hyperparameters, there have been some observations. One of them is the sequential decision process, which means system designers set initial hyperprameters. Then at each step, they decide whether to change them based on the current models, training data at hand, and history of the training process. The other is they decide it based on the performance of the final model found by the supervised ML algorithm, although at each step some immediate reward can be obtained by taking actions. Like previous researches [4], our decision follows the second observation because it results in much smaller final loss that is the long-term rewards. And in the real-world domains, the short-term rewards are not applicable. Based on the recent deep RL for sequential decision problems, we can exploit deep RL techniques and try to learn the Lagrangian the SVM deep neural networks in real-world domains.

We propose an algorithm to learn the Lagrangian multiplier within the actor-critic algorithm [7–10] from deep RL. In particular, an actor network is trained to take an action that decides the Lagrangian multiplier for current step, and a critic network is trained to give feedbacks to the actor network about long-term performance and help the actor network to adjust itself so as to perform better in the future steps [10, 11]. Most of RL studies are based on the weak supervision in the form of a reward given for some of the agent's actions. On the other hand, we consider those of supervised ML with SVM deep neural networks instead of week versions given by a reward. And like previous research [4], with an actor-critic algorithm by learning hyperparameters for the SVM deep neural networks, we propose to feed different training examples to the actor network and the critic network, which improve the generalization performance of the learnt the SVM deep neural networks. By algorithm comparisons, we can show that our algorithm leads to good optimization of Lagrangian multiplier of hyperparameters and can prevent overfitting to a certain extent automatically without human system designers in supervised versions.

2 Materials

2.1 Data

KDDCup 1999 [12] dataset has been widely used for anomaly detection methods. Its training dataset is composed of 4,900,000 single connection vectors that contain 41 features and they are labeled as either normal or an attack. Attack labels (types) fall in one of the following four categories:

- *Denial of Service Attack (DoS)*: This is an attack that intends for making computing or memory resources too busy.
- *User to Root Attach (U2R)*: This attack type starts with access to a normal user account on the system, and then tries to exploit vulnerability to get local access.
- *Remote to Local Attack (R2L)*: This attack occurs when an attacker, who does not have an account on the system, sends packets to the system over a network.
- *Probing*: This is an attempt to gather information about computer networks for the purpose of circumventing its security controls.

3 The Proposed Algorithm

In this section, we propose an actor-critic algorithm that can automate one of SVM hyperparamters, Lagrangian Multiplier for stochastic gradient descent (SGD) based [11] machine learning (ML) algorithms. Our approach is especially for SVM deep neural networks. Most of ML algorithms need to train a model with some parameters ω by minimizing a loss function f defined over a set X of training examples:

$$\omega^* = \mathrm{argmin}_\omega \mathrm{f}_\omega(\mathrm{X}).$$

At every time step, updating parameters by gradients is the standard approach for the loss function minimization called as gradient descent (GD).

$$\omega^{t+1} = \omega^t - \lambda^t \nabla f^t,$$

where λ^t is the learning rate at the step time t, and ∇f^t is the local gradient of f at ω^t.

At a step, we can use the whole batch of the training data, a mini batch of tens/hundreds or a random sample. We can observe that the performance of SGD based methods is quite sensitive to the choice of hyperparamters for non-convex loss function f. Generally, f is usually non-convex with respect to the parameters ω in many ML algorithms, especially for deep neural networks. In terms of SVM deep neural network, we can fine the optimal separating margin by the gradient descent method. Based on the inequality constrained Wolfe Dual form, we can rewrite the SVM as follows [6],

$$\text{Maximize}\left\{\sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i=1}^{l} \alpha_i \alpha_j \mathrm{y}_i \mathrm{y}_j \mathrm{x}_i \mathrm{x}_j\right\},$$

$$C \geqq \alpha_i \geqq 0, \quad i = 1, \ldots, l, \quad \sum_{i=1}^{l} \alpha_i \mathrm{y}_i = 0,$$

where C is a penalty value, α is the Lagrangian multiplier.
We aim to learn the Lagrangian multiplier, one of hyperparameters using RL techniques that can automatically control at each step. Figure 1 shows our automatic Lagrangian multiplier, which adopts the actor-critic framework in RL. The basic idea is that at each step, given the current model ω^t and training sample x, an actor network is used to take an action (the Lagrangian multiplier α_i) and a critic network is used to estimate the goodness of the action. The actor network will be updated using the estimated goodness of α_i, and the critic network will be updated by minimizing temporal difference (TD) Sutton's error [7–10]. We describe the details of our algorithm in the following subsections.

3.1 The Actor

The policy network as an actor network in RL determines the Lagrangian multiplier for the SVM deep neural networks machine learning. There is the proposed algorithm in Fig. 1 based on the current model, training data, and historical information during the training process. Note that ω^t could be of huge dimensions, for example, KDDCup 1999 [12] dataset has been widely used for anomaly detection methods. Its training

```
Algorithm: Actor-Critic Algorithm for Lagrangian multiplier of SVM deep neural networks
Require: Training steps T, training set X, loss function $f$, state function $\chi$, discount factor $\gamma$
Ensure: Model parameters $\omega$, policy parameters $\theta$ of the actor network, value parameters $\rho$ of the critic network, and the action, Lagrangian multiplier $\alpha$
Initial parameters $\omega_0, \theta_0, \rho_0$
for $t$=0, …, $T$ do
    Sample $x_i \in X$, $i \in 1, \ldots, N$ .
    Extract state vector: $s_i^t = \chi(\omega^t, x_i)$.
    Computes Lagrangian multiplier $\alpha_i^t = \pi_\theta(s_i^t), LM(\alpha_i^t)$
    If $LM(\alpha_i^t) > LM(\alpha_{max})$,
        Save $LM(\alpha_i^t)$ as a max,
        Computes the optimized weight vectors, $w=\sum_{i=1}^{N} \alpha_i\, y_i\, x_i$,
        Computes a bias, $b = 1-w^T x_i, y_i=1$
    Compute $\bigtriangledown f^t (x_i)$.
    Update $\omega$: $\omega^{t+1} = \omega^t - \alpha^t_i \bigtriangledown f^t (x_i)$
    Set $r^t = f^t(x_i) - f^{t+1}(x_i)$                    //Update critic network
    Extract state vector: $s^{t+1}{}_i = \chi(\omega^{t+1}, x_i)$.
    Compute $Q\rho\ (s_i^{t+1}, \pi_\theta(s_i^{t+1})), Q\rho(s_i^t, \alpha_i^t)$
    Compute $\delta^t\ = r^t + \gamma Q\rho\ (s_i^{t+1}, \pi_\theta(s_i^{t+1})) - Q\rho(s_i^t, \alpha_i^t)$
    Update $\rho$ by $\nabla\rho = \delta^t \nabla\rho Q\rho(s_i^t, \alpha_i^t)$
    Sample $x_j \in X,\ j \in 1,\ldots,N, j \neq i$          //Update actor network
    Extract state vector: $s_j^{t+1} = \chi(\omega^{t+1}, x_j)$.
    Compute $\alpha_j{}^{t+1} = \pi_\theta(s_j^t)$,
    Update $\theta$ by $\bigtriangledown\theta = \bigtriangledown_{\theta\pi\theta}(s_j^{t+1}) \bigtriangledown_\alpha Q\rho\ (s_j^{t+1}, \alpha_j^{t+1})|\alpha= \pi_\theta(s)$
end for
return
```

Fig. 1 The proposed an actor-critic algorithm for SVM hyperparameters

dataset is composed of 4,900,000 single connection vectors that contain 41 features and they are labeled as either normal or an attack. If the actor network takes all of those parameters, its computational complexity would dominate the complexity of the primary algorithm. Therefore, we propose to use a function $\chi(\cdot)$ to process and yield a compact vector s^t based on SVM deep neural networks as the input of the actor network. In this proposed architecture, the actor learns the Lagrangian multiplier from the on-policy and shows the automatic Lagrangian multiplier. Following the practice in RL, we call $\chi(\cdot)$ the state function, which takes ω^t and the training data x as inputs:

$$s^t = \chi(\omega^t, X):$$

Then the actor network $\pi_\theta(\cdot)$ parameterized by θ yields an action α^t, such as the Lagrangian multiplier:

$$\pi_\theta(s^t) = \alpha^t$$

where the action $\alpha^t \in$ R is a continuous value. When α^{t} is determined, we update the model of the RL. The actor has to learn the input parameters to output a good action. To learn the actor, we need to know how to evaluate the goodness of the actor. The critic network exactly takes this role.

*-8pt

3.2 The Critic

The goal of the proposed algorithm is to find a good policy for the Lagrangian multiplier, α^t to ensure that a good model can be learnt eventually by the SVM deep neural networks. For this purpose, the actor needs to output a good action α^{t} at state s^t so that finally a low training loss $f(\cdot)$ can be achieved. In the value function of deep Q network, $Q_\pi(s, \alpha)$ is often used to denote the long term reward of the state-action pair (s, α) while following the policy π to take future actions such as Lagrangian multiplier. In our proposed algorithm, $Q_\pi(s^t, \alpha^t)$ indicates the accumulative decrement of training loss starting from step t. We define the immediate reward at step t as the one step loss decrement:

$$r^{\mathrm{t}} = f^{\mathrm{t}} - f^{\mathrm{t+1}}.$$

The accumulative value $\mathrm{R}_\pi^{\mathrm{t}}$ of policy π at step t is the total discounted reward from step t:

$$\mathrm{R}_\pi^{\mathrm{t}} = \sum_{(k=t)}^{T} \gamma^{\mathrm{k-t}} r(s^t, \alpha t)$$

where $\gamma \in (0; 1]$ is the discount factor. Given that the states and actions are uncountable in our proposed algorithm, the critic network uses a parametric function $Q\rho(s, \alpha)$ with parameters ρ to approximate the value function of deep Q network, $Q_\pi(s, \alpha)$.
*-8pt

3.3 The Training the Actor and the Critic

The critic has the parameters ρ, which is updated at each step using TD learning [7–10]. More precisely, the critic is trained by minimizing the square error between the estimation $Q\rho(s^t, \alpha^t)$ and the target y^t:

$$y^t = r^t + \gamma Q\rho(s^{t+1}, \alpha^{t+1}) :$$

The TD error is defined as:

$$\begin{aligned}\delta^t &= y^t - \mathrm{Q}\rho(s^t, \alpha^t)\\ &= r^t + \gamma \mathrm{Q}\rho\left(s^{t+1}, \pi_\theta\left(s^{t+1}\right)\right) - \mathrm{Q}\rho(s^t, \alpha^t)\end{aligned}$$

The weight update rule follows the on-policy deterministic actor-critic algorithm. That means that the proposed algorithm learns the Lagrangian multiplier in the way of exploitation. The gradients of critic network are:

$$\nabla\rho = \delta^t \nabla\rho \mathrm{Q}\rho(s^t, \alpha^t)$$

The policy parameters θ of the actor is updated by ensuring that it can output the action with the largest Q value of deep Q network at state s^t, i.e., $\alpha* = \arg\max_\alpha \mathrm{Q}\rho(s^t, \alpha)$. In terms of SVM deep neural networks, the maxim Lagrangian multiplier is that $\alpha* = \arg\max_{\mathrm{LM}(\alpha)} \mathrm{Q}\rho(s^t, \alpha)$. Mathematically,

$$\nabla\theta = \nabla_{\theta\pi\theta}(s^{t+1})\nabla_\alpha \mathrm{Q}\rho(s^{t+1}, \alpha^{t+1})|\alpha = \pi_\theta(\mathrm{s}).$$

3.4 The Algorithm Comparison

In this section, we would like to compare the proposed our algorithm with SVM deep neural networks without automatic Largrangian multiplier. In Fig. 1, there is our algorithm, which is shown that we sample an example, extract the current state vector, and compute Lagrangian multiplier using actor network, update model, update the critic network using TD error, sample another example to update the actor network. In this whole processes are based on on-polity. First, in the proposed algorithm, we consider using only one example for model update like [4]. It is easy to generalize to a mini batch of random examples. That means our proposed whole processes are based on on-polity. There are two categories in learning processes, on-policy and off-policy. Sometimes off-policy is suffering from huge of memories. And second, we use one example (e.g., x_i) for model and the critic network update, but a different example (e.g., x_j) for the actor network update like [4]. Doing so we can avoid that the previous SVM deep neural networks will overfit on some hard examples and can improve the generalization performance of the algorithm on the test set. We can exploit the a-prior knowledge-based advantages by using KDDCup 1999 [12] dataset, which is widely used for anomaly detection with 4 categories based on supervised mode. However, when it comes to neural networks, the network weight are non-convex. If the SVM without neural networks, it can be solvable by quadratic programming because the weight of it is convex. That means we cannot exploit easy example in a classification task. Therefore, we should consider more difficult to be classified correctly with SVM deep neural networks. In the case of hard examples, the gradient descent will be large and the hyperparameters given by the actor network at this step will also be large and intractable. In other words, this hard example will greatly change the model, while it is not a good representative of its category and the hyperparameters should not pay much attention to it. If we input the same example to both the actor and the critic, they will make the model to change a lot to fit the example, while resulting in oscillation of the training. With different example inputs into the actor and critic, it is very likely the critic network will find that the gradient descent direction of the example into the actor is inconsistent with the training example and thus criticize the large Lagrangian multiplier suggested by the actor. More precisely, the update of ω is based on x_i and the Lagrangian multiplier by the actor, while the training target of the actor is to maximize the output of the critic on x_j. If there is big gradient disagreement between x_i and x_j, the update of ω, which is affected by actor's decision, would cause the critic's output on x_j to be small. For optimization, the actor is forced to predict a small Largangian multiplier for a too hard x_i in such a case.

4 Conclusion

We have suggested the algorithm an actor-critic algorithm for SVM hyperparameters. In our proposed algorithm, an actor network is trained to take an action that decides the Lagrangian multiplier for current step, and a critic network is trained to give feedbacks to the actor network about long-term performance and help the actor network to adjust itself so as to perform better in the future steps. We consider those of supervised machine learning (ML) with SVM deep neural networks instead of week versions given by a reward of reinforcement learning. With an actor-critic algorithm by learning hyperparameters for the SVM deep neural networks, we propose to feed different training examples to the actor network and the critic network, which improve the generalization performance of the learnt supervised SVM deep neural networks. By algorithm comparisons, we can find out more difficult to be classified correctly with SVM deep neural networks. In the case of hard examples, the gradient descent will be large and the hyperparameters given by the actor network at this step will also be large and intractable. If we input the different example to both the actor and the critic, they will make the model not to be over fitted with inconsistent training examples. That means we can show that our algorithm leads to good optimization of Lagrangian multiplier of hyperparameters and can prevent overfitting to a certain extent automatically without human system designers in supervised ML algorithms.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2016RIA2B4012386).

References

1. Silver D, Lever G, Heess N (2014) Deterministic policy gradient algorithms. In: ICML'14 proceedings of the 31st international conference on international conference on machine learning, vol 32, pp 387–395
2. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Driessche GVD, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489
3. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. In: NIPS
4. Xu C, Qin T, Wang G, Liu T-Y (2017) Reinforcement learning for learning rate control, under review as a conference paper at ICLR 2017
5. Hansen S (2016) Using deep Q-Learning to control optimization hyperparameters, in ArXiv (2016)
6. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machine. Mach Learn 46:389–422
7. Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3(1):9–44
8. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, vol 1. MIT Press, Cambridge
9. Sutton RS, McAllester DA, Singh SP, Mansour Y (1999) Policy gradient methods for reinforcement learning with function approximation, vol 99. In NIPS, pp 1057—1063
10. Sutton RS (1984) Temporal credit assignment in reinforcement learning. Doctoral Dissertation
11. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. 4(2)
12. Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the KDD CUP 99 dataset. In: Proceedings of the 2009 IEEE symposium on computational intelligence in security and defense applications (CISDA 2009), pp 53–58

Design of the Manufacturing Facility Maintenance Systems Based on Internet of Things

Jin-uk Jung, A. M. Ilham, Min-tae Hwang and Kyo-hong Jin

Abstract There is a strong movement to utilize sensor data which are collected by IoT devices on manufacturing facility. Especially, using these data, companies which manufactures and sells the manufacturing facility or buyers require a kind of maintenance systems for easily managing their facilities because they want to minimize a suspension of work caused by a breakdown on the facility. From their perspective, the suspension means unnecessary expenditure. Therefore, they want to quickly recognize that the breakdown has occurred and restart the facility as soon as possible by solving the problem. In this paper, we design the reference model for the manufacturing facility maintenance systems. The purpose of this systems is to reduce the losses caused by the breakdown.

Keywords Internet of things · Manufacturing facility · Maintenance
Remote control

J. Jung · K. Jin (✉)
Electronic Engineering, Changwon National University, Changwon, South Korea
e-mail: khjin@changwon.ac.kr

J. Jung
e-mail: jaygarcia@changwon.ac.kr

A. M. Ilham
Environment-Friendly Offshore Plant Feed Engineering,
Changwon National University, Changwon, South Korea
e-mail: ilham.anugrah@unpas.ac.id

M. Hwang
Information and Communication Engineering,
Changwon National University, Changwon, South Korea
e-mail: professorhwang@gmail.com

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_65

1 Introduction

Global ICT research firm, Gartner, estimates that global IoT (Internet of Things) device usage will increase by 31% in 2017 to 8.4 billion units in 2017 and will rise to 34% per annum from 2016 to 2020, reaching 20.4 billion units in 2020. The total endpoint and service spending this year is expected to reach $1,699.6 billion, projected to grow at a CAGR of 21% from 2016 to 2020 with a market size of $2.9 trillion by 2020 [1].

In particular, the global market size of industrial IoT is expected to reach $113.7 billion in 2015, and will continue to grow at an annual average rate of 8%, creating a massive market of $195.4 billion by 2022. Industrial IoT is expected to drive the fourth industrial revolution as it applies to all industrial sectors such as transportation, manufacturing, energy, health care, and retailing [2–4].

Since the appearance of Internet of Things, the various kinds of the sensors are installed in the manufacturing facilities of the company to remotely monitor the condition of the facilities and perform the maintenance based on the collected sensor data to prevent the downtime of the facility in advance [5]. This enables companies to reduce the downtime, improve the productivity, and defect rates. The companies selling the manufacturing facility can also create new business models through the remote facility management. For this, we design a reference model for remotely managing the manufacturing facility and explain its operation principle in this paper.

The rest of this paper is organized as follows. We introduce the reference model for the remote manufacturing facility maintenance systems in Sect. 2. In Sects. 3 and 4, we describe about three scenarios with the developed systems and the issues faced by companies that are using systems with our model. Finally, the conclusion and future works follows in Sect. 5.

2 Manufacturing Facility Maintenance Systems

2.1 The Reference Model

Figure 1 is the reference model for Manufacturing Facility Maintenance Systems. This model is composed of six major parts, Manufacturing Facility, Manufacturing Facility Controller, Manufacturing Facility Server Program, Database, Manufacturing Facility Analysis Program, and User Program.

2.2 Manufacturing Facility

We define Manufacturing Facility as a machine which manufactures a product or a part of a product. Since the machine is comprised of many components with own

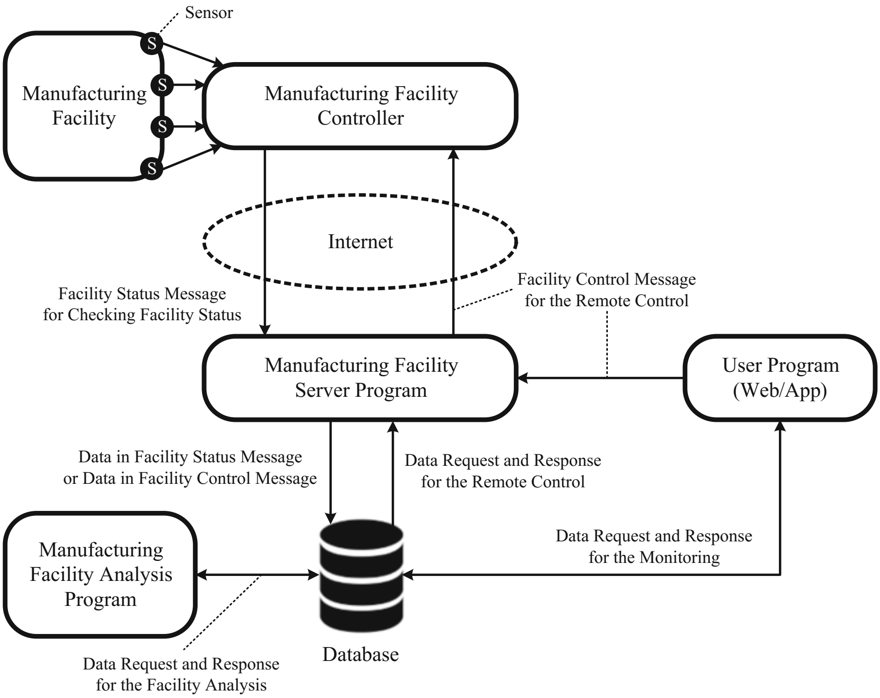

Sensor
S
S
S
S
Manufacturing Facility
Manufacturing Facility Controller
Internet
Facility Status Message for Checking Facility Status
Facility Control Message for the Remote Control
Manufacturing Facility Server Program
User Program (Web/App)
Data in Facility Status Message or Data in Facility Control Message
Data Request and Response for the Remote Control
Manufacturing Facility Analysis Program
Data Request and Response for the Monitoring
Database
Data Request and Response for the Facility Analysis

Fig. 1 The reference model of manufacturing facility maintenance systems

lifetime and is operated by electricity, there will be always a predictive or a non-predictive breakdown. For this reason, sometimes, manufacturers suffer from a halt in production which they are extremely unwilling to face. Therefore, a solution to minimize the downtime, the time period which is wasted by the machine failure, and prevent the breakdown are required.

The solution is to diagnose the facility's condition and take proper actions corresponding to the diagnosis results in real time. To enable this, the various type of sensors must be installed inside or outside the facility, such as a vibration sensor, noise sensor, current sensor, temperature sensor, etc. The designer of the maintenance systems must decide what kind of sensor will be used. Besides, to detect the exact value, the sensors should be installed on and around the critical component. For example, the location which the vibration sensor must be placed on the near the component generating vibrations like a motor. The attachment location of the temperature sensor is around the component radiating heat as the belt for heat transfer. For this, the designer must seriously consider the opinions of the workers who have the valuable experiences and knowledge in dealing with the facility.

The data which the sensors installed in the optimum position detects are continuously transmitted to Manufacturing Facility Controller.

2.3 Manufacturing Facility Controller

Manufacturing Facility Controller is positioned near Manufacturing Facility. The sensors of the facility will be connected to it. The below Fig. 2 is the block diagram of Manufacturing Facility Controller.

Manufacturing Facility Controller has two major roles. The first role is to find a problem or a symptom of Manufacturing Facility with the data from sensors attached in Manufacturing Facility. In Fig. 2, Sensor Data Processing Module receives the data from the sensors and analyzes the data t check the status of Manufacturing Facility. There is a simple method to ascertain the status of Manufacturing Facility. This method is to compare the sensor data with the fixed threshold value. The threshold value can be acquired from a specification of the component or experiences of the skilled experts. If a result of a comparison means an abnormal status, Sensor Data Processing Module activates an alarm or a buzzer to inform the abnormal status of Manufacturing Facility to workers, and then delivers the sensor data to Message Processing Module. Message Processing Module organizes Facility Status Message with the sensor data and transmits it to Manufacturing Facility Server Program.

The second role is to deal with Facility Control Message used for controlling Manufacturing Facility in a remote area. When Message Processing Module receives Facility Control Message, it analyzes the message and sends a command included in this message to Facility Control Module. Facility Control Module controls Manufacturing Facility with a specific operation corresponding to the command, such as power on/off, timer setting, motor speed controlling, etc. However, the most of the small and medium enterprises engaging in the manufacturing industry has poor IT infrastructure on their factory. They don't generally want to pay cost for the use of the static IP addresses due to financial issue. Eventually, Port Forwarding technique [6] with inexpensive wired and wireless router can be a good choice to utilize the remote control function for them.

Meanwhile, the designer must carefully decide a message transmission interval. The most reasonable choice is to use a periodic transmission mixed with an event-driven transmission. That is, an urgent message that must be processed with high priority should be immediately sent using the event-driven transmission. On the other

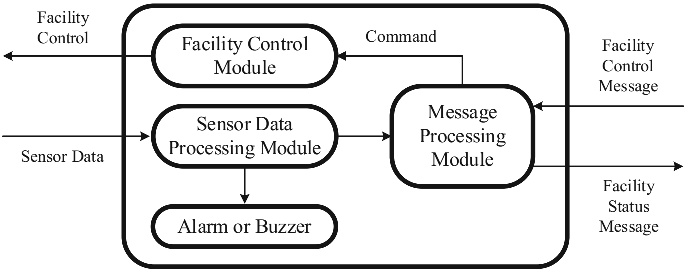


Fig. 2 The block diagram of manufacturing facility controller

hand, a non-urgent message should be sent using the periodic transmission with a fixed time interval. However, if the time interval in the periodic transmission is very short, the event-driven transmission can be ignored.

2.4 Manufacturing Facility Server Program

Manufacturing Facility Server Program operates on a server. This program analyzes the messages and data in the message stores in database. The Fig. 3 illustrates the block diagram of Manufacturing Facility Server Program.

When Facility Status Message is received, the program checks the field marking status of Manufacturing Facility in the message. If the value of this field means the abnormal status, Manufacturing Facility Server Program sends an alarm message to User Program.

Alarm Message Generation Module can be implemented using a notification service like FCM supporting by Google [7]. The Fig. 4 shows a general process that Manufacturing Facility Server Program transmits the alarm message to User Program.

When a user logs or joins in the service web-site with User Program, this program requests a token, a kind of identification to distinguish the user, to FCM server. FCM server receiving the request from User Program replies to the request with the token. This token is unique ID. User Program sends the token to Manufacturing Facility Server Program. The server program stores the token in the database. And if Facility Status Message with the abnormal sensor data is arrived from Manufacturing Facility Controller, the server program creates Alarm Message and delivers to FCM server. Alarm Message is consisted of the contents to be shown on User Program and the token. After find the user corresponding with the token, FCM server sends Alarm Message to the user.

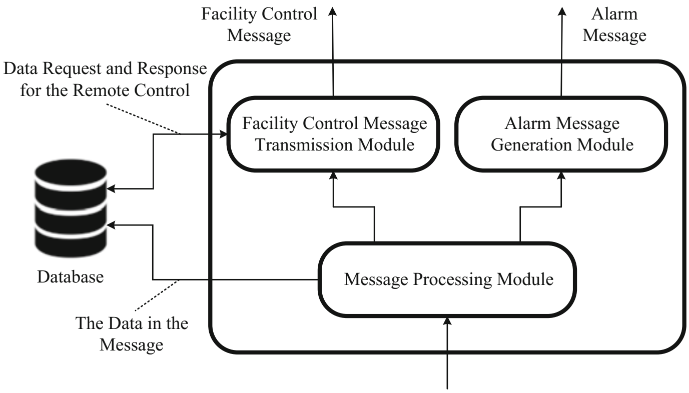

Facility Control Message
Alarm Message
Data Request and Response for the Remote Control
Facility Control Message Transmission Module
Alarm Message Generation Module
Database
Message Processing Module
The Data in the Message

Fig. 3 The block diagram of manufacturing facility server program

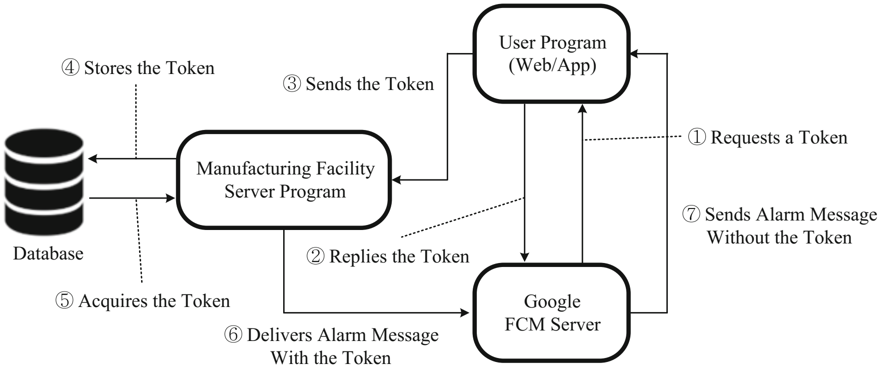


Fig. 4 The process for the transmission of alarm message

Facility Control Message received in Message Processing Module is passed to Facility Control Message Transmission Module. This module acquire the information related with the remote control from Database, such as IP address, port number, etc. With the information, the module sends Facility Control Message to Manufacturing Facility Controller.

2.5 Database

When a user joins in the maintenance service website of Manufacturing Facility through User Program, a database should be automatically created for each company. If not, an administrator for the maintenance service will be suffered from managing the database and tables. In each database, there are five types of tables as shown in Table 1.

Table 1 The tables required for the systems

Table name	Description
Facility data	The significant data in facility status message are stored
Facility information	The general information of each facility in a company are stored, such as facility ID, IP address, port number, etc.
Component lifetime	The lifetime of each component is stored
Failure and repair history	The data related with the failure and repair is stored
Remote control log	In this table, the information that will arbitrate a dispute occurred by using the remote control are stored

Facility Data table is used for storing the significant sensor data in Facility Status Message received from Manufacturing Facility Controller. The data in this table will be used by User Program for the monitoring for the facility.

In Facility Information table, the general information about Manufacturing Facility are stored. This table basically includes facility ID, IP address, and port number, etc., which be utilized for the remote control.

The purpose of Component Lifetime table is to inform the repair or replacement time of a component to the manager of Manufacturing Facility. Since all components of the facility have the own recommended period of use, if the manager of the manufacturing facility misses the time to replace the component, the failure probability of the facility will be increased. Therefore, the manager have to know the appropriate time for the replacement. For this, the repair and replacement time should be managed in this table.

When the breakdown is occurred, the service staff or worker repairs the facility or replaces some parts. The job record which the service staff or worker did is stored in Failure and Repair History table.

If User Program sends the Facility Control Message to Manufacturing Facility Server Program, the program transmits the message to Manufacturing Facility Controller and stores the user command in the message in Remote Control Log table. The remote control is very sensitive function because a field worker may not recognize the control of the remote area. Therefore, a dispute may arise sometimes between a remote user and a field worker. The data recorded in this table will be helpful to arbitrate in the dispute between two sides.

2.6 *Manufacturing Facility Analysis Program*

There are two important factors which affect the lifetime of the facility. These factors are its operating time and operating way. Using the information related with these factors in database, this program predicts the exact lifetime of the component in real time. Even at the same operating time, if the rotation speed of the motor is faster or the torque is large, the lifetime of the equipment may be reduced.

A typical maintenance method is to replace major parts after operating a fixed time based on MTBF (Mean Time Between Failure). However, by periodically checking the condition of the equipment, if the actual replacement time of the parts is checked according to the operation method of the facility, the lifetime of the equipment can be longer. That is, using MTBF with RTBF (Real Time Between Failure) can be a good maintenance method. To apply RTBF, we need to collect data on how the facility is operating and find factors that affect the parts. By using operating steps and various sensor data such as temperature, humidity, rotation speed, power consumption, vibration, noise, etc., it will be possible to make rules that affects the lifetime of the parts and to know the repair time more accurately.

On the other hand, it is difficult to transmit all the data to the server when there is a large amount of data such as vibration and noise among the sensor data. In this case, a method of processing the data just in the facility should also be considered. In other words, edge computing technology should be applied to production facilities, and we will call it Factory Edge which needed more researches.

2.7 User Program

The facility manager or the field-worker, and the selling company of the facility can monitor the condition of the facility and control the facility using User Program in the remote area. For this, there are two kinds of User Program, Web-based User Program and Application for the mobile device called as App.

Web-based User Program is suitable for the manager or the selling company that wants to know every information about the manufacturing facility. Using this program, they can check the condition of the facility though a graph or a table in real time and the information about the facility stored in the Facility Information table or Failure and Repair History table. Besides, they can change the internal setting of the facility or shut the facility down after receiving the alarm message.

The functions of App are identical to Web-Based User Program. However, due to the small screen size of the mobile device, there is the issue for choosing information to be displayed. This issue should be resolved by the selling company because the company exactly knows what the important information of the facility. Also, the field workers who works near the facility can directly know the condition of the facility through the alarm or the buzzer on the facility without the monitoring. Therefore, App should only display the important information which the field-worker need to know.

3 Potential Scenarios

3.1 Facility Status Message Transmission

Figure 5 represents a simple scenario for the transmission of Facility Status Message. First, Manufacturing Facility Controller obtains the data which the sensors attached on Manufacturing Facility detect. The controller analyzes the sensor data and sends Facility Status Message including the analyzed result which represents the status of the facility is normal or abnormal. If the facility is abnormal, it will activate the alarm or buzzer on the facility. Manufacturing Facility Server Program stores the important data extracted in the received Facility Status Message. The stored data is utilized by User Program for the monitoring. If the program finds the abnormal data in Facility Status Message, it will send the alarm message to User Program.

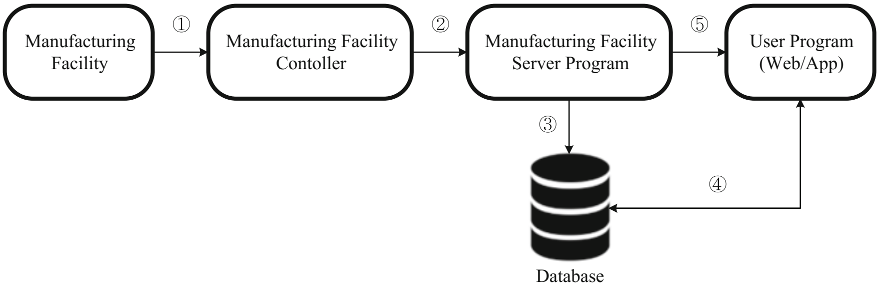

Manufacturing Facility
①
Manufacturing Facility Contoller
②
Manufacturing Facility Server Program
⑤
User Program (Web/App)
③
④
Database

Fig. 5 The scenario for the transmission of facility status message

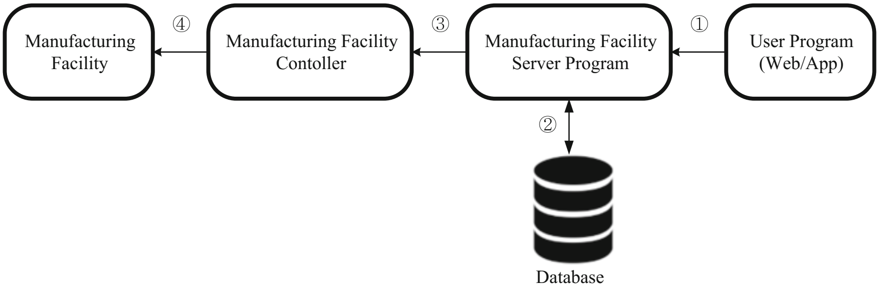

Manufacturing Facility
④
Manufacturing Facility Contoller
③
Manufacturing Facility Server Program
①
User Program (Web/App)
②
Database

Fig. 6 The scenario for the transmission of facility control message

3.2 Facility Control Message Transmission

Figure 6 show the second scenario showing the process of the remote control. First, the manager of the facility writes and sends Facility Control Message in User Program. After receiving this message, Manufacturing Facility Server Program stores this message in the database and acquires IP address and port number of Manufacturing Facility Controller connected to Manufacturing Facility that the manager wants to control from Facility Information table. And then, using these information, the server program sends Facility Control Message to Manufacturing Facility Controller. Manufacturing Facility Controller extracts the certain command for controlling the facility from Facility Control Message and applies the operation for the command to the facility.

3.3 Facility Control Message Transmission

The below Fig. 7 show the scenario for the alarm message transmission. If the value of the sensor data is abnormal, Manufacturing Server Program creates Alarm Message with the token acquired from the database. The token can be found by the id of the

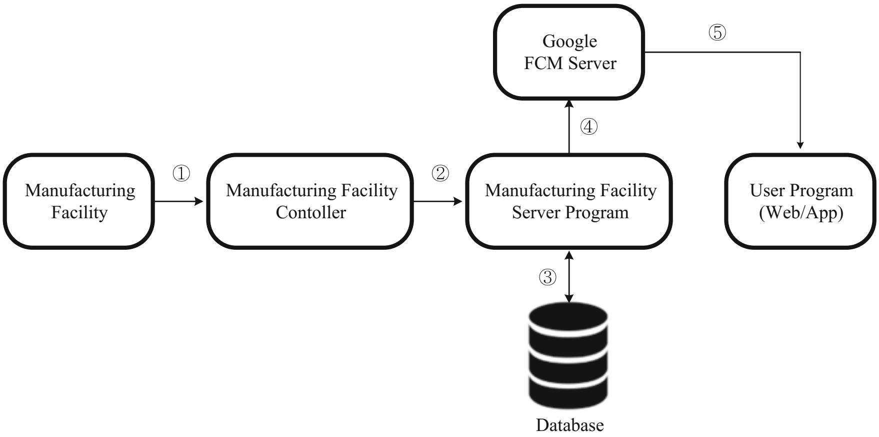


Fig. 7 The scenario for the transmission of alarm message

user possessing the facility with facility ID in Facility Status Message. The server program sends the message to FCM server. FCM server sends Alarm Message to the user who it assigned the token.

4 Issues of Companies Using Systems with the Model

Jung et al. [8] and Kim et al. [9] introduce the systems to which the developed reference model is applied. The reference model was designed based on experiences gained during the development of these systems. The developed systems are being used by relatively poor small and medium enterprises (SMEs) rather than rich big enterprises.

Nowadays, the SMEs using these systems have faced several problems. The most of these issues are related to the maintenance of the system. If the SMEs want to resolve a problem on the system or add a new function, they cannot be deal with it, immediately. This is because most SMEs do not have an administrator who understands and manages the systems due to the financial problem. Even if there is the administrator, this administrator may be not able to respond to all problems.

Besides, these systems are implemented at lowest cost using techniques such as Port Forwarding and FCM mentioned above. Although these techniques can be used at no additional cost, there are disadvantages that Port Forwarding requires troublesome internal setting for access point and FCM notification service does not guarantee accurate and fast message delivery.

In order to solve these issues, eventually, the SMEs needs experts with skills such as hardware design, server management, programming, etc.

5 Conclusion and Future Works

In conclusion, the manufacturer of the manufacturing facility wants to reduce the unnecessary maintenance cost and the purchasing company wants to minimize the downtime spent by the failure of the facility. In this paper, we introduced the manufacturing facility maintenance systems to resolve above two problems. By using this systems, the selling company can efficiently support the maintenance and the customer services in the remote area. The purchasing company can also prevent the suspension of the operation by the unexpected failure. Our future work is to develop new business model based on the manufacturing facility maintenance systems.

Acknowledgements This research was supported by the MIST (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2018-2016-0-00318) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Gartner (2018) Gartner says 8.4 billion connected things will be in use in 2017, up 31 percent from 2016. https://www.gartner.com/newsroom/id/3598917
2. Peter CE, Marco A (2012) Industrial internet pushing the boundaries of minds and machines. Imagination at work
3. World Economic Forum (2015) Accenture: industrial internet of things unleashing the potential of connected products and services
4. McGuinness M (2018) Industrial internet of things will boost economic growth, but greater government and business action needed to fulfill its potential, finds accenture. https://newsroom.accenture.com/news/industrial-internet-of-things-will-boost-economic-growth-but-greater-government-and-business-action-needed-to-fulfill-its-potential-finds-accenture.htm
5. Ramani BV, Amith CA, Jacob MO, Justin B, Thomas P, Vishnu S (2016) Predictive analysis for industrial maintenance automation and optimization using a smart sensor network. In: Internet conference on next generation intelligent systems. IEEE
6. Port forwarding. https://en.wikipedia.org/wiki/Port_forwarding
7. Firebase Cloud Messaging. https://firebase.google.com/products/cloud-messaging/
8. Jung JU, Kim SH, Jin KH (2016) The development of real-time length monitoring systems interlocking with PLC device. In: Proceeding of the 2016 spring conference of Korea institute of information and communication engineering, pp 539–541
9. Kim SH, Jung JM, Jung JU, Jin KH, Hwang MT (2016) Development of the ice machine condition monitoring system for remote diagnosis. In: Proceeding of the 2016 fall conference of Korea institute of information and communication engineering, pp 230–233

Implementation of an Edge Computing Architecture Using OpenStack and Kubernetes

Endah Kristiani, Chao-Tung Yang, Yuan Ting Wang and Chin-Yin Huang

Abstract In the application of the Internet of Things (IoT), all data is stored in the cloud, that causes the long distance of the network logic between the cloud and the device side or client side, this might leads to network delay or slow response time. A challenging issue is how to increase the speed of response time in the cloud computing and the IoT environment for clients. In this paper, we propose a complete set of Edge Computing architecture. There are three layers, namely, Cloud side, Edge side, and Device side. Cloud side mainly deals with more complicated operations and data backup. For overall system infrastructure, we deployed Kubernetes cluster on an OpenStack platform. Edge side optimizes the service of cloud computing systems by performing data processing at the edge of the network. In this phase, we created an Edge Gateway to increase the capacity and performance and reduce the communications bandwidth needed between sensors and the central data.

Keywords Internet of things · Cloud computing · Edge computing · OpenStack Kubernetes

E. Kristiani · C.-T. Yang (✉) · Y. T. Wang
Department of Computer Science, Tunghai University, Taichung 40704, Taiwan, Republic of China
e-mail: ctyang@thu.edu.tw

E. Kristiani
e-mail: endahkristi@gmail.com

Y. T. Wang
e-mail: j8060172@yahoo.com

E. Kristiani · C.-Y. Huang
Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung 40704, Taiwan, Republic of China
e-mail: huangcy@thu.edu.tw

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_66

1 Introduction

As the scale of the IoT gradually expands, topics such as IoT, big data, deep learning, distributed computing and cloud computing have long been inseparable. There is a challenging issue when encountering a long-term accumulation of streaming data or video data (such as Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR)). How to load such a vast amount of data with IoT hardware devices [1]. There is also a discussion about IoT devices, how to combine cloud computing, Big Data processing, and other technologies, how to find the best way to integrate them. These are the topics that we need to explore nowadays [2].

To get the best performance of the IoT application, we integrate open source software to implement a complete set of edge computing architectures [3, 4]. Edge computing between IoT devices and the cloud establishes a relay station to store the data collected by the sensors and provide the most immediate preprocessing and response. We use LoRa, Low Power Wide Area Network (LPWAN), which apply star topology to allow all nodes send data to LoRa Gateway, and then upload to the cloud [5, 6]. In the cloud, OpenStack mainly built as Infrastructure as a Service (IaaS) platform. Through virtualization technology, OpenStack can establish a cluster to provide various virtualized services such as storage virtualization, network virtualization, CPU virtualization and so on [7]. To maximize the utilization of hardware resources, we set up the control node (Kubernetes Master) on OpenStack as a central service control, and use Ceph Storage as a whole system data backup mechanism [8, 9]. On the Edge side, we set up the Raspberry Pi as the Edge Gateway and Kubernetes minion on the Raspberry Pi to provide the service application, which contains the MySQL relational database [10] for data storage.

2 Background Review

2.1 Edge Computing

Edge computing is a method of optimizing cloud computing systems by performing data processing at the edge of the network, near the source of the data, as shown in Fig. 1. IoT applications are appropriate for edge computing architectures. Specifically in emerging of IoT applications such as self-driving, drone, AR/VR, and robotics. All of these recent applications emphasize fast processing capabilities, low latency, and high bandwidth requirements. Therefore, the implementation of edge computing architecture is advantages for the combination of IoT and cloud computing [11].

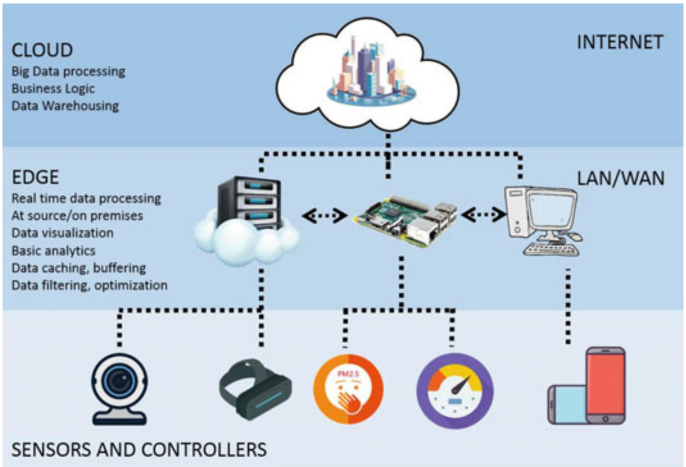


Fig. 1 Edge computing architecture

2.2 OpenStack

OpenStack [12] is a free and open-source software platform for cloud computing, mostly deployed as infrastructure-as-a-service (IaaS), whereby virtual servers and other resources are made available to customers. The software platform consists of interrelated components that control diverse, multi-vendor hardware pools of processing, storage, and networking resources throughout a data center. Users either manage it through a web-based dashboard, through command-line tools or RESTful web services [13]. Figure 2 shows OpenStack architecture.

2.3 Kubernetes

Kubernetes is a robust system, developed by Google, for managing containerized applications in a clustered environment [14]. It aims to provide better ways of handling related, distributed components across the varied infrastructure. Kubernetes also has an ability for managing containerized applications across a cluster of nodes [15].

Kubernetes is particularly well-suited for microservices architectures as shown in Fig. 3. Combining several containers into a single service, Kubernetes also provides a good service discovery mechanism for each service to communicate each other. Most importantly, the modular Kubernetes programming can automatically expand services, not only for rolling updates (Rolling update) and rollback (Rolling back/Undo) but also can integrate CI/CD and other DevOps tools in large-scale containers. Kubernetes is a distributed system which consists of Master as the master used

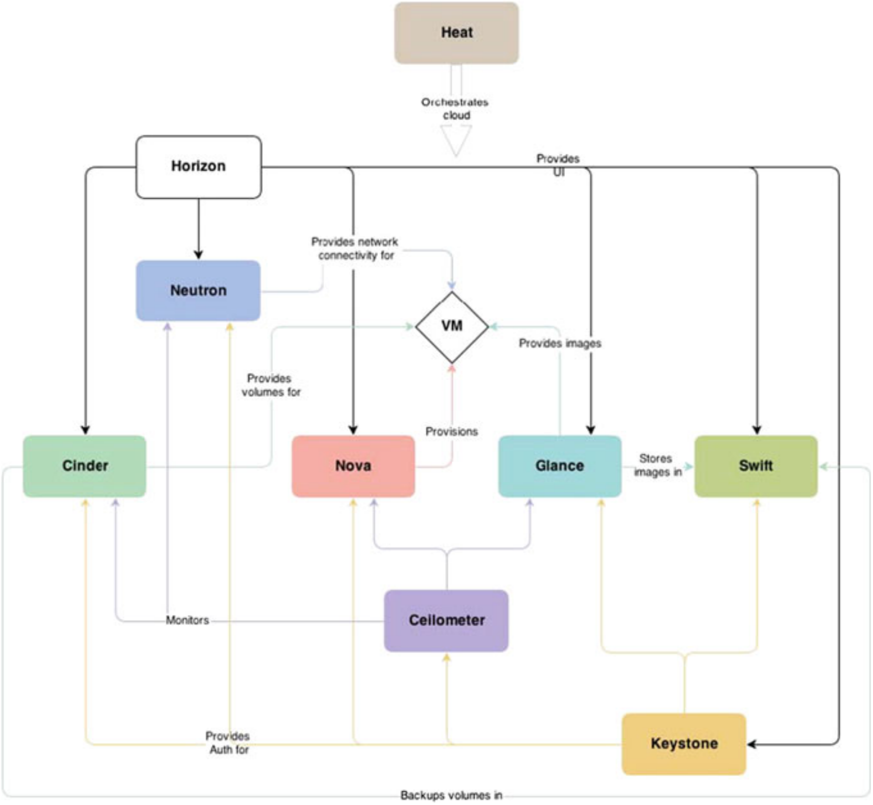


Fig. 2 OpenStack architecture

as the master node and node as a worker, which runs many containers. Kubernetes can handle up to 1,000 nodes or more, uses masters and nodes to deploy cluster. Master contains three basic components, Etcd, API Server, Controller Manager Server. The node consists of four basic components Kubelet, Proxy, Pod, Container. As shown in Fig. 4.

3 System Architecture

To implement Edge Computing, first, we prepared four Lore Node, each node consists of LoRa, and Arduino Uno integrated with one sensor as shown in Fig. 5.

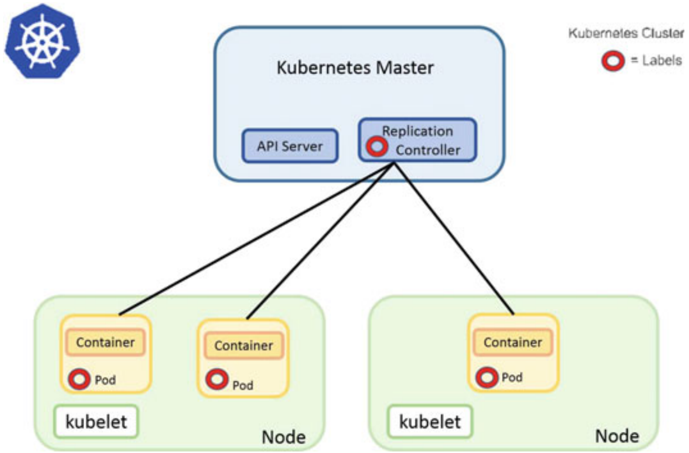


Fig. 3 Kubernetes architecture

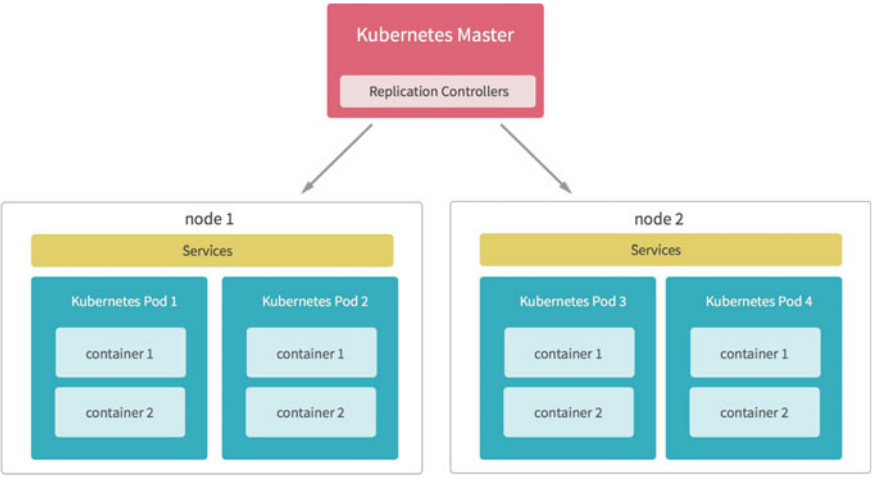


Fig. 4 Kubernetes cluster

Second, we set up one LoRa Gateway as shown in Fig. 6 that has a total of three antennas, designed as two LoRa Shield and a WiFi receiver module. Two sets of LoRa Shield provide dual-band reception, that can bring more stable signal reception. The WiFi module assembled on Arduino, which can serve WiFi data transmission to Edge and Cloud side.

Third, we built a MySQL relational database to serve as sensor data storage for the Edge Gateway. On the Edge Gateway in Fig. 7, we set up a Kubernetes minion to provide services. The service runs on Docker, controlled by Kubernetes, and

Fig. 5 Lora node

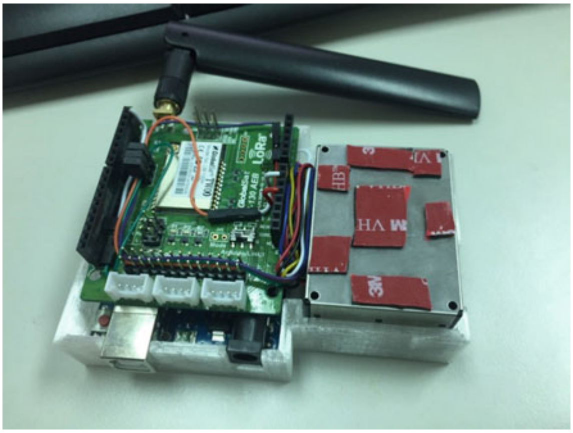
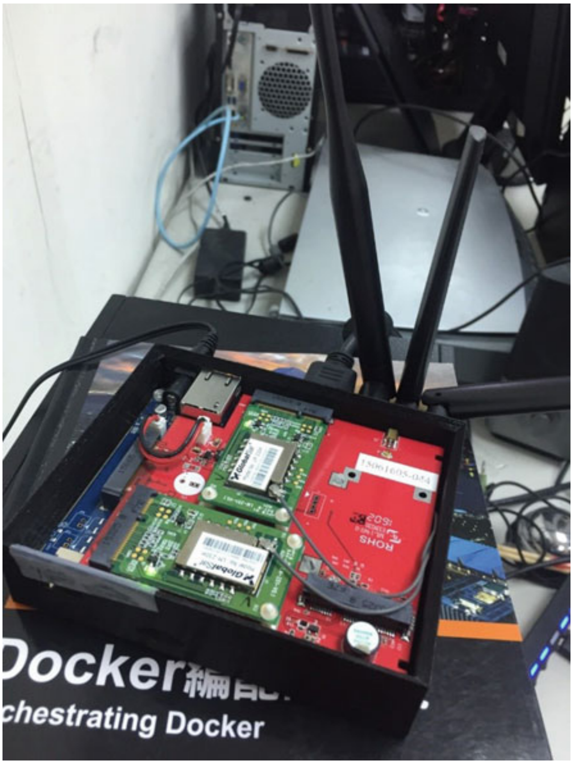
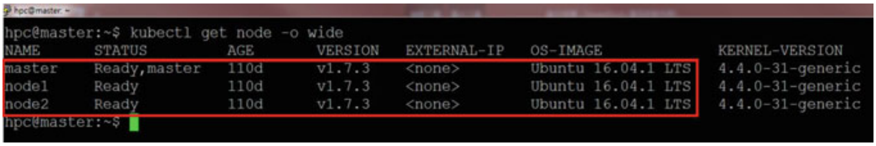


Fig. 6 Lora gateway



Kubernetes is deployed as a clustered version, including one master and two nodes with the operating system using Ubuntu 16.04.1.

On the Cloud side, OpenStack cluster was set up to provide infrastructure services through virtualization technologies. In this study, five nodes were set up as Controller,



```
hpc@master:~$ kubectl get node -o wide
NAME      STATUS          AGE    VERSION   EXTERNAL-IP   OS-IMAGE             KERNEL-VERSION
master    Ready,master    110d   v1.7.3    <none>        Ubuntu 16.04.1 LTS   4.4.0-31-generic
node1     Ready           110d   v1.7.3    <none>        Ubuntu 16.04.1 LTS   4.4.0-31-generic
node2     Ready           110d   v1.7.3    <none>        Ubuntu 16.04.1 LTS   4.4.0-31-generic
hpc@master:~$
```

Fig. 7 Kubernetes cluster

Compute01, Compute02, Block01, and Network. Services established for Nova, Neutron, Keystone, Glance, Cinder, Magnum and Horizon.

Figures 8 and 9 describe overall system architecture that deployed on our system. In this figure, we show scheme system that divided into three parts: Cloud side, Edge side, and Device side.

The Device side contains the sensor and alarm side that can accommodate many different IoT devices and transport agreements. The Cloud side is for the Virtual Machine (VM) and Kubernetes master side and also deal with data backup, complex operations, data visualization and other applications that no need to respond quickly. Finally, the Edge side is mainly responsible for running services and applications, such as data reception and exception notification. Data reception pre-processed via the edge ramp, which significantly reduces the amount of cloud transmission and storage load. These all services run on the Docker container. Besides, the data stored in the edge gateway, as a backup, it also makes closer to the source where the data is collected for reducing latency.

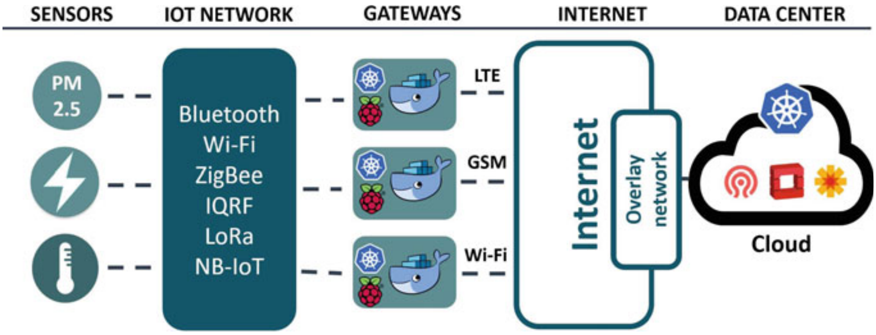


Fig. 8 Edge computing architecture

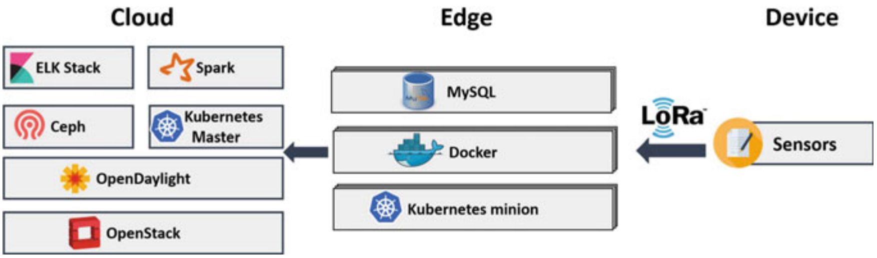


Fig. 9 System architecture

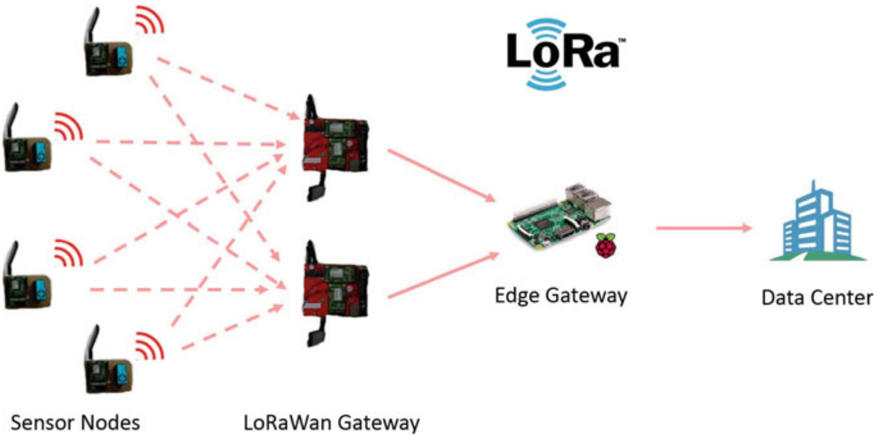


Fig. 10 Device architecture

Figure 10, mainly using Arduino LoRa Shield module and Pms5003t sensor as the overall sensing module and LoRa sensing data collected by LoRa Nodes that transmit to LoRa Gateway through LoRa. In this process, LoRa offers the Low Power Wide Area Network (LPWAN) and the star-topology, and the LoRa Gateway also connected to the Raspberry Pi 3. Finally, the data will be transmitted to the data center (cloud).

4 Experimental Results

LoRa Gateway is placed at the center of the four sensors to provide the most stable data transmission. Generally, LoRa transmission distance around 15–20 km. In this case, we deploy the farthest of two nodes distance about 2 km.

4.1 Booting Time

To test the Pod's start-up time, we boot pods on Kubernetes and create VM on OpenStack. As the number of containers started increases with time, it also found that the number of pods increases. It can be seen from the graph trend line, indicating that many containers are started to run, resulting in delays of the communication and scheduling of the cluster network (Figs. 11 and 12).

Fig. 11 Booting time of kubernetes pods

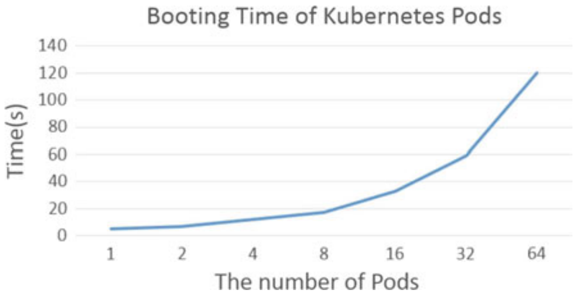
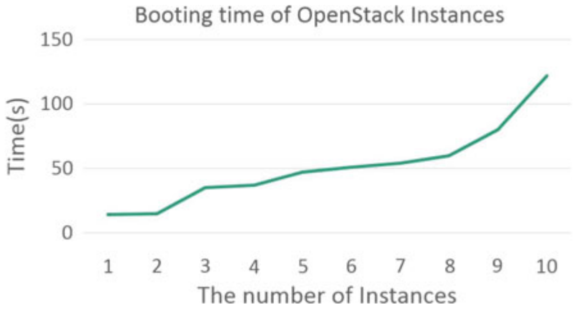


Fig. 12 Booting time of OpenStack instance



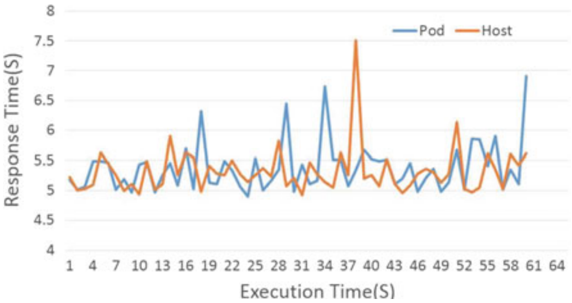


Fig. 13 Network experiment on Kubernetes

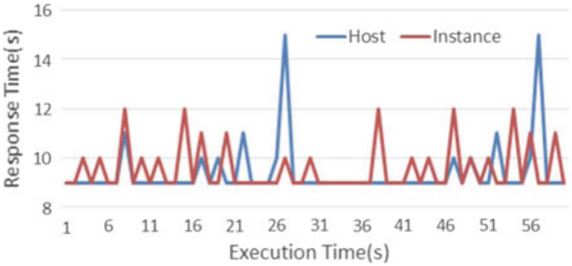


Fig. 14 Network experiment on OpenStack

4.2 Execution Time

A network is essential to edge computing. We tested the network latency between VM on OpenStack and container on Docker, in comparison with the physical machine. We wrote a Dockerfile for ping testing. Then, we execute ping command “ping 8.8.8.8” (Figs. 13 and 14).

5 Conclusion

This paper integrates the technologies of OpenStack, Kubernetes, and Docker to implement a complete set of edge computing architectures. From the experiment, we found that although response time is high sometimes, all range is in 15 ms. This response time is fast and would not make network delay. Therefore, we can know OpenStack can provide the high quality of network performance, and Kubernetes pods also have high network quality like bare metal. The software in our edge computing architecture makes low latency in the network. When we put Raspberry Pi 3 near our applications, we can get fast response time for implementing the real-time applications. In summary, the implementation of Edge Computing Architecture Using OpenStack and Kubernetes poses significant low latency based on our experiment.

Acknowledgements This work was supported in part by the Ministry of Science and Technology, Taiwan R.O.C., under grants number MOST 104-2221-E-029-010-MY3 and MOST 106-3114-E-029-003.

References

1. Cicirelli F, Guerrieri A, Spezzano G, Vinci A (2017) An edge-based platform for dynamic smart city applications. Future Gener Comput Syst
2. Toffetti G, Brunner S, Blöchlinger M, Spillner J, Bohnert TM (2017) Self-managing cloud-native applications: design, implementation, and experience. Future Gener Comput Syst 72:165–179
3. Yang C-T, Chan Y-W, Liu J-C, Lou B-S (2017) An implementation of cloud-based platform with r packages for spatiotemporal analysis of air pollution. J Supercomput, pp 1–22
4. Liu P-Y, Tsan Y-T, Chan Y-W, Chan W-C, Shi Z-Y, Yang C-T, Lou B-S (2018) Associations of pm2. 5 and aspergillosis: ambient fine particulate air pollution and population-based big data linkage analyses. J Ambient Intell Humanized Comput, pp 1–11
5. Yang C-T, Chen S-T, Chang C-H, Den W, Wang Y-T, Kristiani E (2018) Implementation of an intelligent indoor environmental monitoring and management system in cloud. Future Gener Comput Syst
6. Yang C-T, Chen S-T, Chang C-H, Den W, Wu C-C (2018) Implementation of an environmental quality and harmful gases monitoring system in cloud. J Med Biol Eng, pp 1–14
7. Kozhirbayev Z, Richard OS (2017) A performance comparison of container-based technologies for the cloud. Future Gener Comput Syst 68:175–182
8. Yang C-T, Chen C-J, Chen T-Y (2017) Implementation of ceph storage with big data for performance comparison. Lecture notes in electrical engineering, 424:625–633
9. Mikula A, Adamov D, Adam M, Chudoba J, Vec J (2016) Grid site monitoring and log processing using elk, 1787:54–61
10. Shu P, Gu R, Dong Q, Yuan C, Huang Y (2016) Accelerating big data applications on tiered storage system with various eviction policies, pp 1350–1357
11. Giaffreda R, Dupont C, Capra L (2017) Edge computing in iot context: horizontal and vertical linux container migration. In: 2017 Global internet of things summit (GIoTS). IEEE, pp 1–4
12. Openstack (2017). https://www.openstack.org/
13. Yamato Y (2016) Proposal of optimum application deployment technology for heterogeneous iaas cloud, pp 34–37
14. Netto HV, Lung LC, Correia M, Luiz AF, de Souza LMS (2017) State machine replication in containers managed by kubernetes. J Syst Archit 73:53–59
15. Kubernetes (2017) https://kubernetes.io/

Enabling Cross-Domain IoT Interoperability Based on Open Framework

Lei Hang and Do-Hyeun Kim

Abstract IoT platforms are the key solution to provide context data network using the sensor devices, and support backend applications that make sense of the mass of data generated by thousands of sensors. The global IoT platform market continues to rise significantly. To enable the interoperability among heterogeneity of IoT platforms becomes a big challenge for the development trend of the IoT nowadays. This paper presents an open framework based IoT interoperability architecture. This architecture offers the required functionalities for integrating with heterogeneous IoT platforms using open framework based on RESTful. Proposed open framework was designed to facilitate the integration of multiple IoT platforms in different standards. The result of our work indicates that the proposed architecture assists the development of interoperable IoT ecosystems.

Keywords Interoperability · Internet of things · IoT platform · RESTful · IoT ecosystem

1 Introduction

Internet of things (IoT) provides a global connectivity between the real world and a virtual world of entities or things [1]. In this ever-changing landscape of IoT, recent researchers imply that the number of intelligent connected objects will increase dramatically over the next few years [2]. How to manage the ever-increasing number of devices has consistently been a crucial issue in the field of Inter of things. IoT platforms were born to meet this strong demand as a global IoT network always contains millions of devices.

L. Hang (✉) · D.-H. Kim
Computer Engineering Department, Jeju National University, Jeju, South Korea
e-mail: hanglei@jejunu.ac.kr

D.-H. Kim
e-mail: kimdh@jejunu.ac.kr

K. J. Kim and N. Baek (eds.), *Information Science and Applications 2018*,
Lecture Notes in Electrical Engineering 514,
https://doi.org/10.1007/978-981-13-1056-0_67

There are IoT platforms of every shape and size, for specific industries like commercial real estate and family health. Up to now, there were over 300 IoT platforms as of last year and this number will continue to grow at a significant rate of nearly 32% and is expected to reach $1.6 billion market size in 2021 [3]. As a result, there is a pressing demand for multi-domain IoT applications that are in a position to cover multiple fields of the daily routine is increasingly transparent today. At present most of the existing IoT architectures are based on the "closed-loop" concept that focuses on a specific purpose and always being isolated from other organizations and institutions. The connected devices are either implemented within local environments for a definite purpose or integrated with a cloud hosting that is always proprietary.

In this paper, we present an IoT interoperability architecture based on an open cross-layer framework to allow interoperability among heterogeneous cross-domain IoT platforms. This architecture interconnects IoT systems already deployed or new ones added, and support any application field to connect across various IoT domains. This IoT interoperability architecture makes it easy for commercial corporations or personal developers to design networking devices, smart objects, or services, and let them quickly deploy it to the market in order to build new IoT interoperable ecosystems.

The remainder of this paper is organized as follows. Section 2 gives an overview of important aspects of the IoT landscape towards cross-domain IoT interoperability. In the next section, we present the interoperability architecture based on open framework to address the collaboration of cross-domain IoT platforms. Finally, Sect. 4 discusses the conclusion and future work.

2 Related Works

These days, we are coping with varieties of IoT systems, however most of them are either vertically oriented or closed and only open source platforms can be extended quickly to deal with the emergence of new technology. Current IoT platform claims to provide a wide range of solutions for interoperability problems of application developers and is generally open to the creator of third-party applications. We can't apply reusable components or plug-ins to these platforms because of lack of the proprietor.

Many IoT standards have been used to set up the IoT platforms, for example some of them such as CoAP [4] and MQTT [5] are specific for constrained devices with limited computing power and storage, while others like oneM2M [6] and OCF [7] are dedicated to support a universal solution to support a wide variety of backend applications and services. However, these platforms have such a limitation that the communication between devices and platform is under the specific standards. Thus, this causes interoperability issues as developers would wish to create global, cross-platform, and cross-domain applications.

Heterogeneous of IoT standards ultimately prevent the emergence of dynamic IoT ecosystems because the ways for accessing are diverse in terms of the standards which the platforms used. It also produces obstacles to commercial field, especially for small tech enterprise because of a lack of funding to deliver their solution across multiple platforms. They can only provide applications and services for a limited number application scenarios, for example a traffic management system for a specific city.

IoT platform interoperability and federation [8] is a new current solution aims at the specific daily life activities, but limited to the ecological system that can be created around a single centralized platform. Multiple IoT solutions can be set up to provide cross-domain collaboration solutions, sharing of Internet resources through this approach. Such a way would not enforce the implementation of a specific proto-col (network communications, resource representation, and control) as the only standard to in heterogeneous domains [9]. Instead, IoT platforms will have the capability to choose the ideal communication protocols for low-level control and information exchange from the sensors to the IoT gateway to cloud platforms to meet their objectives. Data gathering [10] is a major constraint in IoT while the integration between heterogeneous elements is often in the device or network level. For enabling the end users to communicate the desired data among different platforms and domains, high-level interfaces can be used as the emergence of a new solution called "sensor as a service" [11] can be offered by the domain-specific platforms. With this solution, important and useful information related to a single domain is able to be made available to the third party after preprocessing and aggregation.

3 Proposed IoT Interoperability Architecture Based on Open Framework

Figure 1 overviews the high-level architecture model that enables cross-domain interoperability among heterogeneity of IoT platforms. The proposed architecture is organized around connected devices and IoT platforms within the same space with the open framework. These IoT platforms are separated in different domains with their own IoT solutions focusing on specific communication protocols or activities. The interoperability functionalities provided by the framework are exposed as RESTful APIs so that various devices from different IoT domains can be represented as resources in the network. It also gives developers lots of conveniences when the functionality of the system has an obligation to extend as using the REST technology.

Through the proposed framework, multiple IoT solutions can collaborate so as to provide high-level cross-domain solutions and share IoT resources. Such an approach will not enforce a specific protocol to be established as the standard across different do-mains. On the contrary, IoT platforms will continue to select the desired protocols to control the end-to-end communications and data exchange (from sensors to platforms) that suit their purposes.

The open framework offers high-level APIs for a uniform interface to enable collaboration and support cross-domain discovery and management of IoT resources

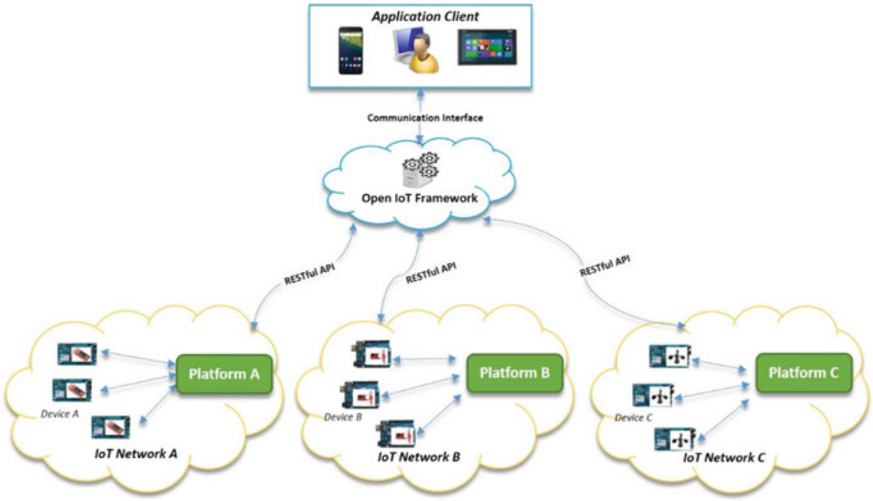


Fig. 1 High-level interoperability architecture model for cross-domain IoT platforms

from diverse platforms. Through this interface, all the interactions can be built around universally supported methods that are responsible for connecting other IoT platforms with the application client. All these platforms can be selected to be cooperative by opening up the access to their resources and by implementing generic high-level APIs. The open framework also provides communication interfaces for different client applications in multiple terminals (Smartphone, tablet, and desktop).

4 Conclusion

IoT surrounds a large number of vertical platforms, each specifically customized by a given scenario and usually adopting proprietary communications, device and resource control. Cross-domain IoT applications which can span wide aspects of daily life is becoming more evident nowadays. This paper overviews the current state of the development of IoT platform market and analyzes the issues that prevent the promotion of global IoT ecosystems. Through the analysis, the IoT interoperability and federation are presented to clear obstacles of the formation of global IoT market. The proposed architecture in this paper has the purpose to ease the implementation of cross-domain applications as well as the integration of smart devices in different IoT environments. We present a high-level approach by using the concept of RESTful so that devices from different domains can be represented as resources in the network regardless of platforms and protocols.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (2014-1-00743) supervised by the IITP (Institute for Information & communications Technology Promotion), and this This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00756, Development of interoperability and management technology of IoT system with heterogeneous ID mechanism). Any correspondence related to this paper should be addressed to DoHyeun Kim; kimdh@jejunu.ac.kr.

References

1. Internet of things. https://en.wikipedia.org/wiki/Internet_of_things. Accessed 15 Nov 2017
2. Gartner Predicts 2015: the internet of things. Accessed 13 Feb 2018
3. Global IoT Platform Market 2017–2021. https://www.technavio.com/report/global-iot-platform-market-2017-2021. Accessed 05 Jan 2018
4. CoAP—Constrained Application Protocol. http://coap.technology/. Accessed 20 Dec 2017
5. MQTT. http://mqtt.org/. Accessed 14 Feb 2018
6. oneM2M Protocol Analysis Technical Report. one M2M-TR-0009, vol. 0, Oct 2013
7. IoTivity project. https://www.iotivity.org. Accessed 13 Dec 2017
8. Swetina J (2014) Toward a standardized common M2M service layer platform: introduction to oneM2M. IEEE Wirel Commun Mag 21(3):20–26
9. Leminen S, Westerlund M, Rajahonka M, Siuruainen R, Andreev S, Balandin S, Koucheryavy Y (2012) Towards IoT ecosystems and business models. Internet of things, smart spaces, and next generation networking. Springer, Berlin, Heidelberg, pp 15–26
10. Soursos S, Podnar-Zarko I, Zwickl P, Gojmerac I, Bianchi G, Carrozzo G (2016) Towards the cross-domain interoperability of IoT platforms. In: 2016 European conference on networks and communication (EUCNC 2016), June 2016
11. Mathew SS, Atif Y, El-Barachi M (2016) From the internet of things to the web of things—enabling by sensing as-a service. In: 2016 12th international conference on innovations in information technology (IIT), Nov 2016